

(19)



Europäisches Patentamt  
European Patent Office  
Office européen des brevets



(11)

EP 0 834 576 B1

(12)

## EUROPEAN PATENT SPECIFICATION

(45) Date of publication and mention  
of the grant of the patent:  
16.01.2002 Bulletin 2002/03

(51) Int Cl.7: C12Q 1/68, G01N 33/566,  
G01N 33/48, C07H 15/12

(21) Application number: 97116548.5

(22) Date of filing: 06.12.1991

### (54) Detection of nucleic acid sequences

Detektion von Nukleinsäuresequenzen

Détection de séquences d'acides nucléiques

(84) Designated Contracting States:  
BE CH DE DK FR GB IT LI NL SE

• Dower, William J.  
Menlo Park, CA 94025 (US)

(30) Priority: 06.12.1990 US 624114

• Solas, Dennis W.  
No. 13 San Francisco, CA 94131 (US)

(43) Date of publication of application:  
08.04.1998 Bulletin 1998/15

(74) Representative: Bizley, Richard Edward et al  
Hepworth, Lawrence, Bryer & Bizley Merlin  
House Falconry Court Baker's Lane  
Epping Essex CM16 5DQ (GB)

(62) Document number(s) of the earlier application(s) in  
accordance with Art. 76 EPC:  
92904971.6 / 0 562 047

(56) References cited:

EP-A- 0 347 210 EP-A- 0 392 546  
WO-A-89/10977 DE-A- 3 722 958

(73) Proprietor: Affymetrix, Inc. (a Delaware  
Corporation)

• KHRAPKO K R ET AL: "AN OLIGONUCLEOTIDE  
HYBRIDIZATION APPROACH TO DNA  
SEQUENCING" FEBS LETTERS, vol. 256, no. 1 -  
02, 9 October 1989, pages 118-122, XP000304574

Santa Clara, CA 95051 (US)

(72) Inventors:

• Fodor, Stephen P.A.  
Palo Alto, CA 94303 (US)

EP 0 834 576 B1

Note: Within nine months from the publication of the mention of the grant of the European patent, any person may give notice to the European Patent Office of opposition to the European patent granted. Notice of opposition shall be filed in a written reasoned statement. It shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

## Description

[0001] The present invention relates to the detection of nucleic acid sequences in two or more collections of nucleic acids.

[0002] The relationship between structure and function of macromolecules is of fundamental importance in the understanding of biological systems. These relationships are important to understanding, for example, the functions of enzymes, structural proteins, and signalling proteins, ways in which cells communicate with each other, as well as mechanisms of cellular control and metabolic feedback.

[0003] Genetic information is critical in continuation of life processes. Life is substantially informationally based and its genetic content controls the growth and reproduction of the organism and its complements. Polypeptides, which are critical features of all living systems, are encoded by the genetic material of the cell. In particular, the properties of enzymes, functional proteins, and structural proteins are determined by the sequence of amino acids which make them up. As structure and function are integrally related, many biological functions may be explained by elucidating the underlying the structural features which provide those functions. For this reason, it has become very important to determine the genetic sequences of nucleotides which encode the enzymes, structural proteins, and other effectors of biological functions. In addition to segments of nucleotides which encode polypeptides, there are many nucleotide sequences which are involved in control and regulation of gene expression.

[0004] The human genome project is directed toward determining the complete sequence the genome of the human organism. Although such a sequence would not correspond to the sequence of any specific individual, it would provide significant information as to the general organization and specific sequences contained within segments from particular individuals. It would also provide mapping information which is very useful for further detailed studies. However, the need for highly rapid, accurate, and inexpensive sequencing technology is nowhere more apparent than in a demanding sequencing project such as this. To complete the sequencing of a human genome would require the determination of approximately  $3 \times 10^9$ , or 3 billion base pairs.

[0005] The procedures typically used today for sequencing include the Sanger dideoxy method, see, e.g., Sanger et al. (1977) *Proc. Natl. Acad. Sci. USA*, 74:5463-5467, or the Maxam and Gilbert method, see, e.g., Maxam et al., (1980) *Methods in Enzymology*, 65:499-559. The Sanger method utilizes enzymatic elongation procedures with chain terminating nucleotides. The Maxam and Gilbert method uses chemical reactions exhibiting specificity of reaction to generate nucleotide specific cleavages. Both methods require a practitioner to perform a large number of complex manual manipulations. These manipulations usually require isolating homogeneous DNA fragments, elaborate and tedious preparing of samples, preparing a separating gel, applying samples to the gel, electrophoresing the samples into this gel, working up the finished gel, and analyzing the results of the procedure.

[0006] The present invention provides a method for detecting nucleic acid sequences in two or more collections of nucleic acids, comprising:

(a) providing an array comprising more than 100 different polynucleotide probes bound to a solid surface;

(b) contacting said array of probes under hybridisation conditions with:

(i) a first collection of nucleic acids comprised of first-labelled nucleic acids having at least some sequences complementary to probes of said array, and

(ii) at least a second collection of nucleic acids comprised of second-labelled nucleic acids having at least some sequences complementary to probes of said array,

wherein said first and second labels are distinguishable from each other; and

(c) detecting hybridisation of first and second labelled complementary nucleic acids to probes of said array.

[0007] In preferred embodiments said first and second labels are fluorescent labels that emit light of different wavelengths.

[0008] The method of the invention may be used to fingerprint at least first and second cells, wherein said first collection of nucleic acids is from a first cell and said second collection of nucleic acids is from a second cell, and fluorescence of said first and second labels hybridised to the array is detected, optionally the method of the invention may further comprise:

(a) determining levels of gene expression in said first and second cells,

(b) determining patterns of gene expression in said first and second cells, or

(c) determining genetic differences between said first and second cells.

[0009] The first and second cells may be different types of cells, optionally wherein:

- (a) at least one cell type is a tumour cell or other cell exhibiting abnormal physiology,
- (b) said first and second cells are at different stages of development,
- (c) said first and second cells are at different stages of infection or other disease, or
- (d) said first and second cells are from different species of organism, optionally wherein said organism is an animal, plant or microorganism.

[0010] In other embodiments at least one collection of nucleic acids may be synthesized by fluorescently labelling:

- (a) RNA isolated, generated or amplified from said cell; or
- (b) DNA isolated, generated or amplified from said cell.

[0011] The solid surface is preferably a polymeric substrate or includes fibers.

[0012] Polynucleotide probes may be bound to the solid surface at a density of at least  $10^3$ , preferably at least  $10^4$ , more preferably at least  $10^5$ , even more preferably at least  $10^6$  regions per  $\text{cm}^2$  to known regions on the solid surface.

[0013] In certain embodiments the solid surface may be formed as a collection of beads and each different polynucleotide probe is bound to a single bead. In such embodiments each bead may further comprise an encoding system bound thereto such that the sequence of the polynucleotide bound to a bead can be determined by decoding the encoding system, optionally wherein said encoding system is selected from the group consisting of a magnetic system, shape encoding system, colour encoding system, or combination thereof. An automated cell sorter may be used to detect hybridisation.

[0014] The array of polynucleotide probes may comprise more than  $10^3$ , preferably more than  $10^4$ , more preferably more than  $10^5$ , even more preferably more than  $10^6$  different probes bound to the solid surface.

[0015] The probes may be greater than about 15, preferably greater than about 25, more preferably greater than about 50 nucleotides in length.

[0016] In certain preferred embodiments at least said two collections of nucleic acids are hybridised to the same array of said probes. The at least said two collections of nucleic acids may be hybridized separately or simultaneously to the same array of said probes.

[0017] The array may be recycled for use.

[0018] The sequences of polynucleotide probes of the array may be known.

[0019] The present invention provides improved methods useful for verification of known sequences, for fingerprinting polymers, and for mapping homologous segments within a sequence. By reducing the number of manual manipulations required and automating most of the steps, the speed, accuracy, and reliability of these procedures are greatly enhanced.

[0020] The production of a substrate having a matrix of positionally defined regions with attached reagents exhibiting known recognition specificity can be used for the sequence analysis of a polymer, fingerprinting, mapping, and general screening of specific interactions.

[0021] The automation of the substrate production method and of the scan and analysis steps minimizes the need for human intervention. This simplifies the tasks and promotes reproducibility.

[0022] The method of the invention employs a composition comprising a plurality of positionally distinguishable sequence specific reagents attached to a solid substrate, which reagents are capable of specifically binding to a predetermined subunit sequence of a preselected multi-subunit length having at least three subunits, said reagents representing substantially all possible sequences of said preselected length. In some embodiments, the subunit sequence is a polynucleotide sequence. In other embodiments, the specific reagent is an oligonucleotide of at least about five nucleotides, preferably at least eight nucleotides, more preferably at least 12 nucleotides. Usually the specific reagents are all attached to a single solid substrate, and the reagents comprise at least 3000 different sequences. In other embodiments, the reagents represent at least about 25% of the possible subsequences of said preselected length. Usually, the reagents are localized in regions of the substrate having a density of at least 25 regions per square centimeter, and often the substrate has a surface area of less than about 4 square centimeters. By way of example and not limitation, fingerprinting methods of the invention may be used for personal identification, genetic screening, identification of pathological conditions, determination of patterns of specific gene expression, and others.

[0023] The detecting of positions which bind target sequence would typically be through a fluorescent label on the target. Although a fluorescent label is probably most convenient, other sorts of labels, e.g., radioactive, enzyme linked, optically detectable, or spectroscopic labels may be used. Because the oligonucleotide probes are positionally defined, the location of the hybridized duplex can directly translate to the sequences which hybridize. Thus analysis of the positions may provide a collection of subsequences found within the target sequence. These subsequences may be matched with respect to their overlaps so as to assemble an intact target sequence.

[0024] Preferred embodiments of the invention will now be described by way of examples and with reference to drawings in which:

[0025] Fig. 1 illustrates a flow chart for sequence, fingerprint, or mapping analysis.

[0026] Fig. 2 illustrates the proper function of a VLSIPS nucleotide synthesis.

[0027] Fig. 3 illustrates the proper function of a VLSIPS dinucleotide synthesis.

[0028] Fig. 4 illustrates the process of a VLSIPS trinucleotide synthesis.

5 I. Overall Description

10 A. general

B. VLSIPS substrates

C. binary masking

D. applications

E. detection methods and apparatus

15 F. data analysis

II. Theoretical Analysis

20 A. simple n-mer structure; theory

B. complications

III. Polynucleotide Sequencing

25 A. preparation of substrate matrix

B. labeling target polynucleotide

C. hybridization conditions

D. detection; VLSIPS scanning

E. analysis

30 F. substrate reuse

IV. Fingerprinting

35 A. general

B. preparation of substrate matrix

C. labeling target nucleotides

D. hybridization conditions

E. detection; VLSIPS scanning

F. analysis

40 G. substrate reuse

H. other polynucleotide aspects

V. Mapping

45 A. general

B. preparation of substrate matrix

C. labeling

D. hybridization/specific interaction

E. detection

F. analysis

50 G. substrate reuse

VI. Additional Screening

55 A. specific interactions

B. sequence comparisons

C. categorizations

D. statistical correlations

## VII. Formation of Substrate

- A. instrumentation
- B. binary masking
- 5 C. synthetic methods
- D. surface immobilization

## VIII. Hybridization/Specific Interaction

- 10 A. general
- B. important parameters

## IX. Detection Methods

- 15 A. labeling techniques
- B. scanning system

## X. Data Analysis

- 20 A. general
- B. hardware
- C. software

## XI. Substrate Reuse

- 25 A. removal of label
- B. storage and preservation
- C. processes to avoid degradation of oligomers

## XII. Integrated Sequencing Strategy

- A. initial mapping strategy
- B. selection of smaller clones

## 35 XIII. Commercial Applications

- A. sequencing
- B. fingerprinting
- C. mapping

## 40 I. OVERALL DESCRIPTION

A. General

45 [0029] The present invention relies in part on the ability to synthesize or attach specific recognition reagents at known locations on a substrate, typically a single substrate. In particular, the present invention provides the ability to prepare a substrate having a very high density matrix pattern of positionally defined specific recognition reagents. The reagents are capable of interacting with their specific targets while attached to the substrate, e.g., solid phase interactions, and by appropriate labeling of these targets, the sites of the interactions between the target and the specific reagents may be derived. Because the reagents are positionally defined, the sites of the interactions will define the specificity of each interaction. As a result, a map of the patterns of interactions with specific reagents on the substrate is convertible into information on the specific interactions taking place, e.g., the recognized features. Where the specific reagents recognize a large number of possible features, this system allows the determination of the combination of specific interactions which exist on the target molecule. Where the number of features is sufficiently large, the identical same combination, or pattern, of features is sufficiently unlikely that a particular target molecule may often be uniquely defined by its features. In the extreme, the features may actually be the subunit sequence of the target molecule, and a given target sequence may be uniquely defined by its combination of features.

50 [0030] The methodology is applicable to sequencing polynucleotides. The specific sequence recognition reagents

will typically be oligonucleotide probes which hybridize with specificity to subsequences found on the target sequence. A sufficiently large number of those probes allows the fingerprinting of a target polynucleotide or the relative mapping of a collection of target polynucleotides, as described in greater detail below.

[0031] In the high resolution fingerprinting provided by a saturating collection of probes which include all possible subsequences of a given size, e.g., 10-mers, collating of all the subsequences and determination of specific overlaps will be derived and the entire sequence can usually be reconstructed.

[0032] Sequence analysis may take the form of complete sequence determination, to the level of the sequence of individual subunits along the entire length of the target sequence. Sequence analysis also may take the form of sequence homology, e.g., less than absolute subunit resolution, where "similarity" in the sequence will be detectable, or the form of selective sequences of homology interspersed at specific or irregular locations.

[0033] In either case, the sequence is determinable at selective resolution or at particular locations. Thus, the hybridization method will be useful as a means for identification, e.g., a "fingerprint", much like a Southern hybridization method is used. It is also useful to map particular target sequences.

## 15 B. VLSIPS Substrates

[0034] The invention is enabled by the development of technology to prepare substrates on which specific reagents may be either positionally attached or synthesized. In particular, the very large scale immobilized polymer synthesis (VLSIPS) technology allows for the very high density production of an enormous diversity of reagents mapped out in a known matrix pattern on a substrate. These reagents specifically recognize subsequences in a target polymer and bind thereto, producing a map of positionally defined regions of interaction. These map positions are convertible into actual features recognized, and thus would be present in the target molecule of interest.

[0035] As indicated, the sequence specific recognition reagents will often be oligonucleotides which hybridize with fidelity and discrimination to the target sequence.

[0036] In the generic sense, the VLSIPS technology allows the production of a substrate with a high density matrix of positionally mapped regions with specific recognition reagents attached at each distinct region. By use of protective groups which can be positionally removed, or added, the regions can be activated or deactivated for addition of particular reagents or compounds. Details of the protection are described below and in PCT publication no. WO90/15070, published December 13, 1990. In a preferred embodiment, photosensitive protecting agents will be used and the processes of activation or deactivation may be controlled by electro-optical and optical methods, similar to many of the processes used in semiconductor wafer and chip fabrication.

[0037] In the nucleic acid nucleotide sequencing application, a VLSIPS substrate is synthesized having positionally defined oligonucleotide probes. See PCT publication no. WO90/15070, published December 13, 1990. By use of masking technology and photosensitive synthetic subunits, the VLSIPS apparatus allows for the stepwise synthesis of polymers according to a positionally defined matrix pattern. Each oligonucleotide probe will be synthesized at known and defined positional locations on the substrate. This forms a matrix pattern of known relationship between position and specificity of interaction. The VLSIPS technology allows the production of a very large number of different oligonucleotide probes to be simultaneously and automatically synthesized including numbers in excess of about  $10^2$ ,  $10^3$ ,  $10^4$ ,  $10^5$ ,  $10^6$ , or even more, and at densities of at least about  $10^2$ ,  $10^3/\text{cm}^2$ ,  $10^4/\text{cm}^2$ ,  $10^5/\text{cm}^2$  and up to  $10^6/\text{cm}^2$  or more. This application discloses methods for synthesizing polymers on a silicon or other suitably derivatized substrate, methods and chemistry for synthesizing specific types of biological polymers on those substrates, apparatus for scanning and detecting whether interaction has occurred at specific locations on the substrate, and various other technologies related to the use of a high density very large scale immobilized polymer substrate. In particular, sequencing, fingerprinting, and mapping applications are discussed herein in detail, though related technologies are described WO 91/17271 (PCT/US91/02989).

[0038] The regions which define particular reagents will usually be generated by selective protecting groups which may be activated or deactivated. Typically the protecting group will be bound to a monomer subunit or spatial region, and can be spatially affected by an activator, such as electromagnetic radiation. Examples of protective groups with utility herein include nitroveratryl oxycarbonyl (NVOC), nitrobenzyl oxycarbonyl (NBOC) or  $\alpha,\alpha$ -dimethyl-dimethoxybenzyl oxycarbonyl (DEZ).

## C. Binary Masking

[0039] There are various particular ways to optimize the synthetic processes.

[0040] Briefly, the binary synthesis strategy refers to an ordered strategy for parallel synthesis of diverse polymer sequences by sequential addition of reagents which may be represented by a reactant matrix, and a switch matrix, the product of which is a product matrix. A reactant matrix is a  $1 \times n$  matrix of the building blocks to be added. The switch matrix is all or a subset of the binary numbers from 1 to  $n$  arranged in columns. In preferred embodiments, a binary

strategy is one in which at least two successive steps illuminate half of a region of interest on the substrate. In most preferred embodiments, binary synthesis refers to a synthesis strategy which also factors a previous addition step. For example, a strategy in which a switch matrix for a masking strategy halves regions that were previously illuminated, illuminating about half of the previously illuminated region and protecting the remaining half (while also protecting about half of previously protected regions and illuminating about half of previously protected regions). It will be recognized that binary rounds may be interspersed with non-binary rounds and that only a portion of a substrate may be subjected to a binary scheme, but will still be considered to be a binary masking scheme within the definition herein. A binary "masking" strategy is a binary synthesis which uses light to remove protective groups from materials for addition of other materials such as nucleotides.

5 [0041] In particular, this procedure provides a simplified and highly efficient method for saturating all possible sequences of a defined length polymer. This masking strategy is also particularly useful in producing all possible oligonucleotide sequence probes of a given length.

#### D. Applications

10 [0042] The technology provided by the present invention has very broad applications. Although described specifically for polynucleotide sequences, similar sequencing, fingerprinting, mapping, and screening procedures may be applied to polypeptide, carbohydrate, or other polymers. This may be for *de novo* sequencing, or may be used in conjunction with a second sequencing procedure to provide independent verification. See, e.g., (1988) *Science* 242:1245. For example, a large polynucleotide sequence defined by either the Maxam and Gilbert technique or by the Sanger technique may be verified by using the present invention.

15 [0043] In addition, by selection of appropriate probes, a polynucleotide sequence can be fingerprinted. Fingerprinting is a less detailed sequence analysis which usually involves the characterization of a sequence by a combination of defined features. Sequence fingerprinting is particularly useful because the repertoire of possible features which can be tested is virtually infinite. Moreover, the stringency of matching is also variable depending upon the application. A Southern Blot analysis may be characterized as a means of simple fingerprint analysis.

20 [0044] Fingerprinting analysis may be performed to the resolution of specific nucleotides, or may be used to determine homologies, most commonly for large segments. In particular, an array of oligonucleotide probes of virtually any workable size may be positionally localized on a matrix and used to probe a sequence for either absolute complementary matching, or homology to the desired level of stringency using selected hybridization conditions.

25 [0045] In addition, the present invention provides means for mapping analysis of a target sequence or sequences. Mapping will usually involve the sequential ordering of a plurality of various sequences, or may involve the localization of a particular sequence within a plurality of sequences. This may be achieved by immobilizing particular large segments onto the matrix and probing with a shorter sequence to determine which of the large sequences contain that smaller sequence. Alternatively, relatively shorter probes of known or random sequence may be immobilized to the matrix and a map of various different target sequences may be determined from overlaps. Principles of such an approach are described in some detail by Evans et al. (1989) "Physical Mapping of Complex Genomes by Cosmid Multiplex Analysis," *Proc. Natl. Acad. Sci. USA* 86:5030-5034; Michiels et al. (1987) "Molecular Approaches to Genome Analysis: A Strategy for the Construction of Ordered Overlap Clone Libraries," *CABIOS* 3:203-210; Olsen et al. (1986) "Random-Clone Strategy for Genomic Restriction Mapping in Yeast," *Proc. Natl. Acad. Sci. USA* 83:7826-7830; Craig, et al. (1990) "Ordering of Cosmid Clones Covering the Herpes Simplex Virus Type I (HSV-I) Genome: A Test Case for Fingerprinting by Hybridization," *Nuc. Acids Res.* 18:2653-2660; and Coulson, et al. (1986) "Toward a Physical Map of the Genome of the Nematode *Caenorhabditis elegans*," *Proc. Natl. Acad. Sci. USA* 83:7821-7825.

30 [0046] Fingerprinting analysis also provides a means of identification. In addition to its value in apprehension of criminals from whom a biological sample, e.g., blood, has been collected, fingerprinting can ensure personal identification for other reasons. For example, it may be useful for identification of bodies in tragedies such as fire, flood, and vehicle crashes. In other cases the identification may be useful in identification of persons suffering from amnesia, or of missing persons. Other forensics applications include establishing the identity of a person, e.g., military identification "dog tags", or may be used in identifying the source of particular biological samples. Fingerprinting technology is described, e.g., in Carrano, et al. (1989) "A High-Resolution, Fluorescence-Based, Semi-automated method for DNA Fingerprinting," *Genomics* 4: 129-136. See, e.g., table I, for nucleic acid applications.

TABLE I

VLSIPS PROJECT IN NUCLEIC ACIDS		
5	I. Construction of Chips	
10	II. Applications	
15	A.	Sequencing
20	1.	Primary sequencing
25	2.	Secondary sequencing (sequence checking)
30	3.	Large scale mapping
35	4.	Fingerprinting
40	B.	Duplex/Triplex formation
45	1.	Antisense
50	2.	Sequence specific function modulation (e.g. promoter inhibition)
55	C.	Diagnosis
60	1.	Genetic markers
65	2.	Type markers
70	a.	Blood donors
75	b.	Tissue transplants
80	D.	Microbiology
85	1.	Clinical microbiology
90	2.	Food microbiology
95	III.	Instrumentation
100	A.	Chip machines
105	B.	Detection
110	IV.	Software Development
115	A.	Instrumentation software
120	B.	Data reduction software
125	C.	Sequence analysis software

[0047] The fingerprinting analysis may be used to perform various types of genetic screening. For example, a single substrate may be generated with a plurality of screening probes, allowing for the simultaneous genetic screening for a large number of genetic markers. Thus, prenatal or diagnostic screening can be simplified, economized, and made more generally accessible.

[0048] In addition to the sequencing, fingerprinting, and mapping applications, the present invention also provides means for determining specificity of interaction with particular sequences.

#### E. Detection Methods and Apparatus

[0049] An appropriate detection method applicable to the selected labeling method can be selected. Suitable labels include radionucleotides, enzymes, substrates, cofactors, inhibitors, magnetic particles, heavy metal atoms, and particularly fluorescers, chemiluminescers, and spectroscopic labels. Patents teaching the use of such labels include U. S. Patent Nos. 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149; and 4,366,241.

[0050] With an appropriate label selected, the detection system best adapted for high resolution and high sensitivity detection may be selected. As indicated above, an optically detectable system, e.g., fluorescence or chemiluminescence would be preferred. Other detection systems may be adapted to the purpose, e.g., electron microscopy, scanning electron microscopy (SEM), scanning tunneling electron microscopy (STEM), infrared microscopy, atomic force microscopy (AFM), electrical conductance, and image plate transfer.

[0051] With a detection method selected, an apparatus for scanning the substrate will be designed. Apparatus, as described in PCT publication no. WO90/15070, published December 13, 1990, is particularly appropriate. Design modifications may also be incorporated therein.

5 F. Data Analysis

[0052] Data is analyzed by processes similar to those described below in the section describing theoretical analysis. More efficient algorithms will be mathematically devised, and will usually be designed to be performed on a computer. Various computer programs which may more quickly or efficiently make measurement samples and distinguish signal 10 from noise will also be devised.

[0053] The initial data resulting from the detection system is an array of data indicative of fluorescent intensity versus location on the substrate. The data are typically taken over regions substantially smaller than the area in which synthesis of a given polymer has taken place. Merely by way of example, if polymers were synthesized in squares on the substrate having dimensions of 500 microns by 500 microns, the data may be taken over regions having dimensions of 5 microns 15 by 5 microns. In most preferred embodiments, the regions over which fluorescence data are taken across the substrate are less than about 1/2 the area of the regions in which individual polymers are synthesized, preferably less than 1/10 the area in which a single polymer is synthesized, and most preferably less than 1/100 the area in which a single polymer is synthesized. Hence, within any area in which a given polymer has been synthesized, a large number of fluorescence data points are collected.

[0054] A plot of number of pixels versus intensity for a scan should bear a rough resemblance to a bell curve, but spurious data are observed, particularly at higher intensities. Since it is desirable to use an average of fluorescent intensity over a given synthesis region in determining relative binding affinity, these spurious data will tend to undesirably skew the data.

[0055] Accordingly, in one embodiment of the invention the data are corrected for removal of these spurious data 25 points, and an average of the data points is thereafter utilized in determining relative binding efficiency. In general the data are fitted to a base curve and statistically measures are used to remove spurious data.

[0056] In an additional analytical tool, various degeneracy reducing analogues may be incorporated in the hybridization probes. Various aspects of this strategy are described, e.g., in Macevicz, S. (1990) PCT publication number WO 90/04652.

30 II. THEORETICAL ANALYSIS

[0057] The principle of the hybridization sequencing procedure is based, in part, upon the ability to determine overlaps 35 of short segments. The VLSIPS technology provides the ability to generate reagents which will saturate the possible short subsequence recognition possibilities. The principle is most easily illustrated by using a binary sequence, such as a sequence of zeros and ones. Once having illustrated the application to a binary alphabet, the principle may easily be understood to encompass three letter, four letter, five or more letter, even 20 letter alphabets. A theoretical treatment 40 of analysis of subsequence information to reconstruction of a target sequence is provided, e.g., in Lysov, Yu., et al. (1988) *Doklady Akademii Nauk. SSR* 303:1508-1511; Khropko K., et al. (1989) *FEBS Letters* 256:118-122; Pevzner, P. (1989) *J. of Biomolecular Structure and Dynamics* 7:63-69; and Drmanac, R. et al. (1989) *Genomics* 4:114-128.

[0058] The reagents for recognizing the subsequences will usually be specific for recognizing a particular polymer 45 subsequence anywhere within a target polymer. It is preferable that conditions may be devised which allow absolute discrimination between high fidelity matching and very low levels of mismatching. The reagent interaction will preferably exhibit no sensitivity to flanking sequences, to the subsequence position within the target, or to any other remote structure within the sequence.

A. Simple n-mer Structure: Theory

50 1. Simple two letter alphabet: example

[0059] A simple example is presented below of how a sequence of ten digits comprising zeros and ones would be sequenceable using short segments of five digits. For example, consider the sample ten digit sequence:

1010011100. A VLSIPS substrate could be constructed, as discussed elsewhere, which would have reagents attached in a defined matrix pattern which specifically recognize each of the possible five digit sequences of ones and zeros. The number of possible five digit subsequences is  $2^5 = 32$ . The number of possible different sequences 10 digits long is  $2^{10} = 1,024$ . The five contiguous digit subsequences within a ten digit sequence number six, i.e., positioned at digits 1-5, 2-6, 3-7, 4-8, 5-9, and 6-10. It will be noted that the specific order of the digits in the sequence is important and that the order is directional, e.g., running left to right versus right to left. The first five digit sequence contained in

the target sequence is 10100. The second is 01001, the third is 10011, the fourth is 00111, the fifth is 01110, and the sixth is 11100.

[0060] The VLSIPS substrate would have a matrix pattern of positionally attached reagents which recognize each of the different 5-mer subsequences. Those reagents which recognize each of the 6 contained 5-mers will bind the target, and a label allows the positional determination of where the sequence specific interaction has occurred. By correlation of the position in the matrix pattern, the corresponding bound subsequences can be determined.

[0061] In the above-mentioned sequence, six different 5-mer sequences would be determined to be present. They would be:

10

10100  
01001  
10011  
00111  
01110  
11100

15

[0062] Any sequence which contains the first five digit sequence, 10100, already narrows the number of possible sequences (e.g., from 1024 possible sequences) which contain it to less than about 192 possible sequences.

[0063] This 192 is derived from the observation that with the subsequence 10100 at the far left of the sequence, in positions 1-5, there are only 32 possible sequences. Likewise, for that particular subsequence in positions 2-6, 3-7, 4-8, 5-9, and 6-10. So, to sum up all of the sequences that could contain 10100, there are 32 for each position and 6 positions for a total of about 192 possible sequences. However, some of these 10 digit sequences will have been counted twice. Thus, by virtue of containing the 10100 subsequence, the number of possible 10-mer sequences has been decreased from 1024 sequences to less than about 192 sequences.

[0064] In this example, not only do we know that sequence contains 10100, but we also know that it contains the second five character sequence, 01001. By virtue of knowing that the sequence contains 10100, we can look specifically to determine whether the sequence contains a subsequence of five characters which contains the four leftmost digits plus a next digit to the left. For example, we would look for a sequence of X1010, but we find that there is none. Thus, we know that the 10100 must be at the left end of the 10-mer. We would also look to see whether the sequence contains the rightmost four digits plus a next digit to the right, e.g., 0100X. We find that the sequence also contains the sequence 01001, and that X is a 1. Thus, we know at least that our target sequence has an overlap of 0100 and has the left terminal sequence 101001.

[0065] Applying the same procedure to the second 5-mer, we also know that the sequence must include a sequence of five digits having the sequence 1001Y where Y must be either 0 or 1. We look through the fragments and we see that we have a 10011 sequence within our target, thus Y is also 1. Thus, we would know that our sequence has a sequence of the first seven being 1010011.

[0066] Moving to the next 5-mer, we know that there must be a sequence of 0011Z, where Z must be either 0 or 1. We look at the fragments produced above and see that the target sequence contains a 00111 subsequence and Z is 1. Thus, we know the sequence must start with 10100111.

[0067] The next 5-mer must be of the sequence 0111W where W must be 0 or 1. Again, looking up at the fragments produced, we see that the target sequence contains a 01110 subsequence, and W is a 0. Thus, our sequence to this point is 101001110. We know that the last 5-mer must be either 11100 or 11101. Looking above, we see that it is 11100 and that must be the last of our sequence. Thus, we have determined that our sequence must have been 1010011100.

[0068] However, it will be recognized from the example above with the sequences provided therein, that the sequence analysis can start with any known positive probe subsequence. The determination may be performed by moving linearly along the sequence checking the known sequence with a limited number of next positions. Given this possibility, the sequence may be determined, besides by scanning all possible oligonucleotide probe positions, by specifically looking only where the next possible positions would be. This may increase the complexity of the scanning but may provide a longer time span dedicated towards scanning and detecting specific positions of interest relative to other sequence possibilities. Thus, the scanning apparatus could be set up to work its way along a sequence from a given contained oligonucleotide to only look at those positions on the substrate which are expected to have a positive signal.

[0069] It is seen that given a sequence, it can be de-constructed into n-mers to produce a set of internal contiguous subsequences. From any given target sequence, we would be able to determine what fragments would result. The hybridization sequence method depends, in part, upon being able to work in the reverse, from a set of fragments of known sequences to the full sequence. In simple cases, one is able to start at a single position and work in either or both directions towards the ends of the sequence as illustrated in the example.

[0070] The number of possible sequences of a given length increases very quickly with the length of that sequence.

Thus, a 10-mer of zeros and ones has 1024 possibilities, a 12-mer has 4096. A 20-mer has over a million possibilities, and a 30-mer has over a billion. However, a given 30-mer has, at most, 26 different internal 5-mer sequences. Thus, a 30 character target sequence having over a million possible sequences can be substantially defined by only 26 different 5-mers. It will be recognized that the probe oligonucleotides will preferably, but need not necessarily, be of identical length, and that the probe sequences need not necessarily be contiguous in that the overlapping subsequences need not differ by only a single subunit. Moreover, each position of the matrix pattern need not be homogeneous, but may actually contain a plurality of probes of known sequence. In addition, although all of the possible subsequence specifications would be preferred, a less than full set of sequences specifications could be used. In particular, although a substantial fraction will preferably be at least about 70%, it may be less than that. About 20% would be preferred, 10 more preferably at least about 30% would be desired. Higher percentages would be especially preferred.

## 2. Example of four letter alphabet

[0071] A four letter alphabet may be conceptualized in at least two different ways from the two letter alphabet. One way, is to consider the four possible values at each position and to analogize in a similar fashion to the binary example each of the overlaps. A second way is to group the binary digits into groups.

[0072] Using the first means, the overlap comparisons are performed with a four letter alphabet rather than a two letter alphabet. Then, in contrast to the binary system with 10 positions where  $2^{10} = 1024$  possible sequences, in a 4-character alphabet with 10 positions, there will actually be  $4^{10} = 1,048,576$  possible sequences. Thus, the complexity of a four character sequence has a much larger number of possible sequences compared to a two character sequence. Note, however, that there are still only 6 different internal 5-mers. For simplicity, we shall examine a 5 character string with 3 character subsequences. Instead of only 1 and 0, the characters may be designated, e.g., A, C, G, and T. Let us take the sequence GGCTA. The 3-mer subsequences are:

25

GGC  
GCT  
CTA

30

Given these subsequences, there is one sequence, or at most only a few sequences which would produce that combination of subsequences, i.e., GGCTA.

[0073] Alternatively, with a four character universe, the binary system can be looked at in pairs of digits. The pairs would be 00, 01, 10, and 11. In this manner, the earlier used sequence 1010011100 is looked at as 10,10,01,11,00. Then the first character of two digits is selected from the possible universe of the four representations 00, 01, 10, and 11. Then a probe would be in an even number of digits, e.g., not five digits, but, three pairs of digits or six digits. A similar comparison is performed and the possible overlaps determined. The 3-pair subsequences are:

40

10,10,01  
10,01,11  
01,11,00

45

and the overlap reconstruction produces 10,10,01,11,00.

50

[0074] The latter of the two conceptual views of the 4 letter alphabet provides a representation which is similar to what would be provided in a digital computer. The applicability to a four nucleotide alphabet is easily seen by assigning, e.g., 00 to A, 01 to C, 10 to G, and 11 to T. And, in fact, if such a correspondence is used, both examples for the 4 character sequences can be seen to represent the same target sequence. The applicability of the hybridization method and its analysis for determining the ultimate sequence is easily seen if A is the representation of adenine, C is the representation of cytosine, G is the representation of guanine, and T is the representation of thymine or uracil.

## B. Complications

[0075] Two obvious complications exist with the method of sequence analysis by hybridization. The first results from a probe of inappropriate length while the second relates to internally repeated sequences.

[0076] The first obvious complication is a problem which arises from an inappropriate length of recognition sequence, which causes problems with the specificity of recognition. For example, if the recognized sequence is too short, every sequence which is utilized will be recognized by every probe sequence. This occurs, e.g., in a binary system where

the probes are each of sequences which occur relatively frequently, e.g., a two character probe for the binary system. Each possible two character probe would be expected to appear  $\frac{1}{4}$  of the time in every single two character position. Thus, the above sequence example would be recognized by each of the 00, 10, 01, and 11. Thus, the sequence information is virtually lost because the resolution is too low and each recognition reagent specifically binds at multiple sites on the target sequence.

[0077] The number of different probes which bind to a target depends on the relationship between the probe length and the target length. At the extreme of short probe length, the just mentioned problem exists of excessive redundancy and lack of resolution. The lack of stability in recognition will also be a problem with extremely short probes. At the extreme of long probe length, each entire probe sequence is on a different position of a substrate. However, a problem arises from the number of possible sequences, which goes up dramatically with the length of the sequence. Also, the specificity of recognition begins to decrease as the contribution to binding by any particular subunit may become sufficiently low that the system fails to distinguish the fidelity of recognition. Mismatched hybridization may be a problem with the polynucleotide sequencing applications, though the fingerprinting and mapping applications may not be so strict in their fidelity requirements. As indicated above, a thirty position binary sequence has over a million possible sequences, a number which starts to become unreasonably large in its required number of different sequences, even though the target length is still very short. Preparing a substrate with all sequence possibilities for a long target may be extremely difficult due to the many different oligomers which must be synthesized.

[0078] The above example illustrates how a long target sequence may be reconstructed with a reasonably small number of shorter subsequences. Since the present day resolution of the regions of the substrate having defined oligomer probes attached to the substrate approaches about 10 microns by 10 microns for resolvable regions, about  $10^6$ , or 1 million, positions can be placed on a one centimeter square substrate. However, high resolution systems may have particular disadvantages which may be outweighed using the lower density substrate matrix pattern. For this reason, a sufficiently large number of probe sequences can be utilized so that any given target sequence may be determined by hybridization to a relatively small number of probes.

[0079] A second complication relates to convergence of sequences to a single subsequence. This will occur when a particular subsequence is repeated in the target sequence. This problem can be addressed in at least two different ways. The first, and simpler way, is to separate the repeat sequences onto two different targets. Thus, each single target will not have the repeated sequence and can be analyzed to its end. This solution, however, complicates the analysis by requiring that some means for cutting at a site between the repeats can be located. Typically a careful sequencer would want to have two intermediate cut points so that the intermediate region can also be sequenced in both directions across each of the cut points. This problem is inherent in the hybridization method for sequencing but can be minimized by using a longer known probe sequence so that the frequency of probe repeats is decreased.

[0080] Knowing the sequence of flanking sequences of the a repeat will simplify the use of polymerase chain reaction (PCR) or a similar technique to further definitively determine the sequence between sequence repeats. Probes can be made to hybridize to those known sequences adjacent the repeat sequences, thereby producing new target sequences for analysis. See, e.g., Innis et al. (eds.) (1990) PCR Protocols: A Guide to Methods and Applications, Academic Press; and methods for synthesis of oligonucleotide probes, see, e.g., Gait (1984) Oligonucleotide Synthesis: A Practical Approach, IRL Press, Oxford.

[0081] Other means for dealing with convergence problems include using particular longer probes, and using degeneracy reducing analogues, see, e.g., Macevicz, S. (1990) PCT publication number WO 90/04652. By use of stretches of the degeneracy reducing analogues with other probes in particular combinations, the number of probes necessary to fully saturate the possible oligomer probes is decreased. For example, with a stretch of 12-mers having the central 4-mer of degenerate nucleotides, in combination with all of the possible 8-mers, the collection numbers twice the number of possible 8-mers, e.g.  $65,536 + 65,536 = 131,072$ , but the population provides screening equivalent to all possible 12-mers.

[0082] By way of further explanation, all possible oligonucleotide 8-mers may be depicted in the fashion:  
N1-N2-N3-N4-N5-N6-N7-N8,

in which there are  $4^8 = 65,536$  possible 8-mers. Producing all possible 8-mers requires  $4 \times 8 = 32$  chemical binary synthesis steps to produce the entire matrix pattern of 65,536 8-mer possibilities. By incorporating degeneracy reducing nucleotides, D's, which hybridize nonselectively to any corresponding complementary nucleotide, new oligonucleotides 12-mers can be made in the fashion:

N1-N2-N3-N4-D-D-N5-N6-N7-N8,

in which there are again, as above, only  $4^8 = 65,536$  possible "12-mers", which in reality only have 8 different nucleotides.

[0083] However, it can be seen that each possible 12-mer probe could be represented by a group of the two 8-mer types. Moreover, repeats of less than 12 nucleotides would not converge, or cause repeat problems in the analysis. Thus, instead of requiring a collection of probes corresponding to all 12-mers, or  $4^{12} = 16,777,216$  different 12-mers, the same information can be derived by making 2 sets of "8-mers" consisting of the typical 8-mer collection of  $4^8 =$

65,536 and the "12-mer" set with the degeneracy reducing analogues, also requiring making  $4^8 = 65,536$ . The combination of the two sets, requires making  $65,536 + 65,536 = 131,072$  different molecules, but giving the information of 16,777,216 molecules. Thus, incorporating the degeneracy reducing analogue decreases the number of molecules necessary to get 12-mer resolution by a factor of about 128-fold.

5

### III. POLYNUCLEOTIDE SEQUENCING

[0084] In principle, the making of a substrate having a positionally defined matrix pattern of all possible oligonucleotides of a given length involves a conceptually simple method of synthesizing each and every different possible oligonucleotide, and affixed to a definable position. Oligonucleotide synthesis is presently mechanized and enabled by current technology and instruments supplied by Applied Biosystems, Foster City, California.

10

#### A. Preparation of Substrate Matrix

15

[0085] The production of the collection of specific oligonucleotides used in polynucleotide sequencing may be produced in at least two different ways. Present technology certainly allows production of ten nucleotide oligomers on a solid phase or other synthesizing system. See, e.g., instrumentation provided by Applied Biosystems, Foster City, California. Although a single oligonucleotide can be relatively easily made, a large collection of them would typically require a fairly large amount of time and investment. For example, there are  $4^{10} = 1,048,576$  possible ten nucleotide oligomers. Present technology allows making each and every one of them in a separate purified form though such might be costly and laborious.

20

[0086] Once the desired repertoire of possible oligomer sequences of a given length have been synthesized, this collection of reagents may be individually positionally attached to a substrate, thereby allowing a batchwise hybridization step. Present technology also would allow the possibility of attaching each and every one of these 10-mers to a separate specific position on a solid matrix. This attachment could be automated in any of a number of ways, particularly use of a caged biotin type linking. This would produce a matrix having each of different possible 10-mers.

25

[0087] A batchwise hybridization is much preferred because of its reproducibility and simplicity. An automated process of attaching various reagents to positionally defined sites on a substrate is provided in PCT publication no. WO90/15070, and PCT publication no. WO91/07087.

30

[0088] Instead of separate synthesis of each oligonucleotide, these oligonucleotides are conveniently synthesized in parallel by sequential synthetic processes on a defined matrix pattern as provided in PCT publication no. WO90/15070. Here, the oligonucleotides are synthesized stepwise on a substrate at positionally separate and defined positions. Use of photosensitive blocking reagents allows for defined sequences of synthetic steps over the surface of a matrix pattern. By use of the binary masking strategy, the surface of the substrate can be positioned to generate a desired pattern of regions, each having a defined sequence oligonucleotide synthesized and immobilized thereto.

35

[0089] Although the prior art technology can be used to generate the desired repertoire of oligonucleotide probes, an efficient and cost effective means would be to use the VLSIPS technology described in PCT publication no. WO90/15070. In this embodiment, the photosensitive reagents involved in the production of such a matrix are described below.

40

[0090] The regions for synthesis may be very small, usually less than about  $100 \mu\text{m} \times 100 \mu\text{m}$ , more usually less than about  $50 \mu\text{m} \times 50 \mu\text{m}$ . The photolithography technology allows synthetic regions of less than about  $10 \mu\text{m} \times 10 \mu\text{m}$ , about  $3 \mu\text{m} \times 3 \mu\text{m}$ , or less. The detection also may detect such sized regions, though larger areas are more easily and reliably measured.

45

[0091] At a size of about 30 microns by 30 microns, one million regions would take about 11 centimeters square or a single wafer of about 4 centimeters by 4 centimeters. Thus the present technology provides for making a single matrix of that size having all one million plus possible oligonucleotides. Region size are sufficiently small to correspond to densities of at least about 5 regions/cm<sup>2</sup>, 20 regions/cm<sup>2</sup>, 50 regions/cm<sup>2</sup>, 100 regions/cm<sup>2</sup>, and greater, including 300 regions/cm<sup>2</sup>, 1000 regions/cm<sup>2</sup>, 3K regions/cm<sup>2</sup>, 10K regions/cm<sup>2</sup>, 30K regions/cm<sup>2</sup>, 100K regions/cm<sup>2</sup>, 300K regions/cm<sup>2</sup> or more, even in excess of one million regions/cm<sup>2</sup>.

50

[0092] Although the pattern of the regions which contain specific sequences is theoretically not important, for practical reasons certain patterns will be preferred in synthesizing the oligonucleotides. Binary masking algorithms can be applied to generate the pattern of known oligonucleotide probes.

55

[0093] By use of these binary masks, a highly efficient means is provided for producing the substrate with the desired matrix pattern of different sequences. Although the binary masking strategy allows for the synthesis of all lengths of polymers, the strategy may be easily modified to provide only polymers of a given length. This is achieved by omitting steps where a subunit is not attached.

55

[0094] The strategy for generating a specific pattern may take any of a number of different approaches. However, the binary masking and binary synthesis approaches provide a maximum of diversity with a minimum number of actual

synthetic steps.

[0095] The length of oligonucleotides used in sequencing applications will be selected on criteria determined to some extent by the practical limits discussed above. For example, if probes are made as oligonucleotides, there will be 65,536 possible eight nucleotide sequences. If a nine subunit oligonucleotide is selected, there are 262,144 possible permutations of sequences. If a ten-mer oligonucleotide is selected, there are 1,048,576 possible permutations of sequences. As the number gets larger, the required number of positionally defined subunits necessary to saturate the possibilities also increases. With respect to hybridization conditions, the length of the matching necessary to converse stability of the conditions selected can be compensated for. See, e.g., Kanehisa, M. (1984) *Nuc. Acids Res.* 12:203-213.

[0096] Although not described in detail here, but below for oligonucleotide probes, the VLSIPS technology would typically use a photosensitive protective group on an oligonucleotide. Sample oligonucleotides are shown in Figure 4. In particular, the photoprotective group on the nucleotide molecules may be selected from a wide variety of positive light reactive groups preferably including nitro aromatic compounds such as o-nitrobenzyl derivatives or benzylsulfonyl. See, e.g., Gait (1984) *Oligonucleotide Synthesis: A Practical Approach*, IRL Press, Oxford. In a preferred embodiment, 6-nitro-veratryl oxycarbonyl (NVOC), 2-nitrobenzyl oxycarbonyl (NBOC), or  $\alpha, \alpha$ -dimethyl-dimethoxybenzyl oxycarbonyl (DEZ) is used. Photoremovable protective groups are described in, e.g., Patchornik (1970) *J. Amer. Chem. Soc.* 92: 6333; and Amit et al. (1974) *J. Organic Chem.* 39:192.

[0097] A preferred linker is used to attach the oligonucleotide to a silicon matrix. A more detailed description is provided below. A photosensitive blocked nucleotide may be attached to specific locations of unblocked prior cycles of attachments on the substrate and can be successively built up to the correct length oligonucleotide probe.

[0098] It should be noted that multiple substrates may be simultaneously exposed to a single target sequence where each substrate is a duplicate of one another or where, in combination, multiple substrates together provide the complete or desired subset of possible subsequences. This provides the opportunity to overcome a limitation of the density of positions on a single substrate by using multiple substrates. In the extreme case, each probe might be attached to a single bead or substrate and the beads sorted by whether there is a binding interaction. Those beads which do bind might be encoded to indicate the subsequence specificity of reagents attached thereto.

[0099] Then, the target may be bound to the whole collection of beads and those beads that have appropriate specific reagents on them will bind to target. Then a sorting system may be utilized to sort those beads that actually bind the target from those that do not. This may be accomplished by presently available cell sorting devices or a similar apparatus. After the relatively small number of beads which have bound the target have been collected, the encoding scheme may be read off to determine the specificity of the reagent on the bead. An encoding system may include a magnetic system, a shape encoding system, a color encoding system, or a combination of any of these, or any other encoding system. Once again, with the collection of specific interactions that have occurred, the binding may be analyzed for sequence information, fingerprint information, or mapping information.

[0100] The parameters of polynucleotide sizes of both the probes and target sequences are determined by the applications and other circumstances. The length of the oligonucleotide probes used will depend in part upon the limitations of the VLSIPS technology to provide the number of desired probes. For example, in an absolute sequencing application, it is often useful to have virtually all of the possible oligonucleotides of a given length. As indicated above, there are 65,536 8-mers, 262,144 9-mers, 1,048,576 10-mers, 4,194,304 11-mers, etc. As the length of the oligomer increases the number of different probes which must be synthesized also increases at a rate of a factor of 4 for every additional nucleotide. Eventually the size of the matrix and the limitations in the resolution of regions in the matrix will reach the point where an increase in number of probes becomes disadvantageous. However, this sequencing procedure requires that the system be able to distinguish, by appropriate selection of hybridization and washing conditions, between binding of absolute fidelity and binding of complementary sequences containing mismatches. On the other hand, if the fidelity is unnecessary, this discrimination is also unnecessary and a significantly longer probe may be used. Significantly longer probes would typically be useful in fingerprinting or mapping applications.

[0101] The length of the probe is selected for a length that it will bind with specificity to possible targets. The hybridization conditions are also very important in that they will determine how close the homology of complementary binding will be detected. In fact, a single target may be evaluated at a number of different conditions to determine its spectrum of specificity for binding particular probes. This may find use in a number of other applications besides the polynucleotide sequencing fingerprinting or mapping. In a related fashion, different regions with reagents having differing affinities or levels of specificity may allow such a spectrum to be defined using a single incubation, where various regions, at a given hybridization condition, show the binding affinity. For example, fingerprint probes of various lengths, or with specific defined non-matches may be used. Unnatural nucleotides or nucleotides exhibiting modified specificity of complementary binding are described in greater detail in Macevicz (1990) PCT pub. No. WO 90/04652; and see the section on modified nucleotides in the Sigma Chemical Company catalogue.

B. Labeling Target Nucleotide

[0102] The label used to detect the target sequences will be determined, in part, by the detection methods being applied. Thus, the labeling method and label used are selected in combination with the actual detecting systems being used.

[0103] Once a particular label has been selected, appropriate labeling protocols will be applied, as described below for specific embodiments. Standard labeling protocols for nucleic acids are described, e.g., in Sambrook et al.; Kambara, H. et al. (1988) *BioTechnology* 6:816-821; Smith, L. et al. (1985) *Nuc. Acids Res.* 13:2399-2412; for polypeptides, see, e.g., Allen G. (1989) *Sequencing of Proteins and Peptides*, Elsevier, New York, especially chapter 5, and Greenstein and Winitz (1961) *Chemistry of the Amino Acids*, Wiley and Sons, New York. Carbohydrate labeling is described, e.g., in Chaplin and Kennedy (1986) *Carbohydrate Analysis: A Practical Approach*, IRL Press, Oxford. Labeling of other polymers will be performed by methods applicable to them as recognized by a person having ordinary skill in manipulating the corresponding polymer.

[0104] In some embodiments, the target need not actually be labeled if a means for detecting where interaction takes place is available. As described below, for a nucleic acid embodiment, such may be provided by an intercalating dye which intercalates only into double stranded segments, e.g., where interaction occurs. See, e.g., Sheldon et al. U.S. Pat. No. 4,582,789.

[0105] In many uses, the target sequence will be absolutely homogeneous, both with respect to the total sequence and with respect to the ends of each molecule. Homogeneity with respect to sequence is important to avoid ambiguity. It is preferable that the target sequences of interest not be contaminated with a significant amount of labeled contaminating sequences. The extent of allowable contamination will depend on the sensitivity of the detection system and the inherent signal to noise of the system. Homogeneous contamination sequences will be particularly disruptive of the sequencing procedure.

[0106] However, although the target polynucleotide must have a unique sequence, the target molecules need not have identical ends. In fact, the homogeneous target molecule preparation may be randomly sheared to increase the numerical number of molecules. Since the total information content remains the same, the shearing results only in a higher number of distinct sequences which may be labeled and bind to the probe. This fragmentation may give a vastly superior signal relative to a preparation of the target molecules having homogeneous ends. The signal for the hybridization is likely to be dependent on the numerical frequency of the target-probe interactions. If a sequence is individually found on a larger number of separate molecules a better signal will result. In fact, shearing a homogeneous preparation of the target may often be preferred before the labeling procedure is performed, thereby producing a large number of labeling groups associated with each subsequence.

C. Hybridization Conditions

[0107] The hybridization conditions between probe and target should be selected such that the specific recognition interaction, i.e., hybridization, of the two molecules is both sufficiently specific and sufficiently stable. See, e.g., Hames and Higgins (1985) *Nucleic Acid Hybridisation: A Practical Approach*, IRL Press, Oxford. These conditions will be dependent both on the specific sequence and often on the guanine and cytosine (GC) content of the complementary hybrid strands. The conditions may often be selected to be universally equally stable independent of the specific sequences involved. This typically will make use of a reagent such as an arylammonium buffer. See, Wood et al. (1985) "Base Composition-independent Hybridization in Tetramethylammonium Chloride: A Method for Oligonucleotide Screening of Highly Complex Gene Libraries," *Proc. Natl. Acad. Sci. USA*, 82:1585-1588; and Krupov et al. (1989) "An Oligonucleotide Hybridization Approach to DNA Sequencing," *FEBS Letters*, 256:118-122. An arylammonium buffer tends to minimize differences in hybridization rate and stability due to GC content. By virtue of the fact that sequences then hybridize with approximately equal affinity and stability, there is relatively little bias in strength or kinetics of binding for particular sequences. Temperature and salt conditions along with other buffer parameters should be selected such that the kinetics of renaturation should be essentially independent of the specific target subsequence or oligonucleotide probe involved. In order to ensure this, the hybridization reactions will usually be performed in a single incubation of all the substrate matrices together exposed to the identical same target probe solution under the same conditions.

[0108] Alternatively, various substrates may be individually treated differently. Different substrates may be produced, each having reagents which bind to target subsequences with substantially identical stabilities and kinetics of hybridization. For example, all of the high GC content probes could be synthesized on a single substrate which is treated accordingly. In this embodiment, the arylammonium buffers could be unnecessary. Each substrate is then treated in a manner that the collection of substrates show essentially uniform binding and the hybridization data of target binding to the individual substrate matrix is combined with the data from other substrates to derive the necessary subsequence binding information. The hybridization conditions will usually be selected to be sufficiently specific that the fidelity of base matching will be properly discriminated. Of course, control hybridizations should be included to determine the

stringency and kinetics of hybridization.

**D. Detection; VLSIPS Scanning**

5 [0109] The next step of the sequencing process by hybridization involves labeling of target polynucleotide molecules. A quickly and easily detectable signal is preferred. The VLSIPS apparatus is designed to easily detect a fluorescent label, so fluorescent tagging of the target sequence is preferred. Other suitable labels include heavy metal labels, magnetic probes, chromogenic labels (e.g., phosphorescent labels, dyes, and fluorophores) spectroscopic labels, enzyme linked labels, radioactive labels, and labeled binding proteins. Additional labels are described in U.S. Pat. No. 10 4,366,241.

10 [0110] The detection methods used to determine where hybridization has taken place will typically depend upon the label selected above. Thus, for a fluorescent label a fluorescent detection step will typically be used. PCT publication no. WO90/15070 describes apparatus and mechanisms for scanning a substrate matrix using fluorescence detection, but a similar apparatus is adaptable for other optically detectable labels.

15 [0111] The detection method provides a positional localization of the region where hybridization has taken place. However, the position is correlated with the specific sequence of the probe since the probe has specifically been attached or synthesized at a defined substrate matrix position. Having collected all of the data indicating the subsequences present in the target sequence, this data may be aligned by overlap to reconstruct the entire sequence of the target, as illustrated above.

20 [0112] It is also possible to dispense with actual labeling if some means for detecting the positions of interaction between the sequence specific reagent and the target molecule are available. This may take the form of an additional reagent which can indicate the sites either of interaction, or the sites of lack of interaction, e.g., a negative label. For the nucleic acid embodiments, locations of double strand interaction may be detected by the incorporation of intercalating dyes, or other reagents such as antibody or other reagents that recognize helix formation, see, e.g., Sheldon, et al. (1986) U.S. Pat. No. 4,582,789.

**E. Analysis**

30 [0113] Although the reconstruction can be performed manually as illustrated above, a computer program will typically be used to perform the overlap analysis. A program may be written and run on any of a large number of different computer hardware systems. The variety of operating systems and languages useable will be recognized by a computer software engineer. Various different languages may be used, e.g., BASIC; C; PASCAL; etc. A simple flow chart of data analysis is illustrated in Figure 1.

35 **F. Substrate Reuse**

40 [0114] Finally, after a particular sequence has been hybridized and the pattern of hybridization analyzed, the matrix substrate should be reusable and readily prepared for exposure to a second or subsequent target polynucleotides. In order to do so, the hybrid duplexes are disrupted and the matrix treated in a way which removes all traces of the original target. The matrix may be treated with various detergents or solvents to which the substrate, the oligonucleotide probes, and the linkages to the substrate are inert. This treatment may include an elevated temperature treatment, treatment with organic or inorganic solvents, modifications in pH, and other means for disrupting specific interaction. Thereafter, a second target may actually be applied to the recycled matrix and analyzed as before.

45 **IV. FINGERPRINTING**

**A. General**

50 [0115] Many of the procedures and techniques used in the polynucleotide sequencing section are also appropriate for fingerprinting applications. See, e.g., Poustka, et al. (1986) Cold Spring Harbor Symposia on Quant. Biol., vol. LI, 131-139, Cold Spring Harbor Press, New York. The fingerprinting method provided herein is based, in part, upon the ability to positionally localize a large number of different specific probes onto a single substrate. This high density matrix pattern provides the ability to screen for, or detect, a very large number of different sequences simultaneously. In fact, depending upon the hybridization conditions, fingerprinting to the resolution of virtually absolute matching of sequence is possible thereby approaching an absolute sequencing embodiment. And the sequencing embodiment is very useful in identifying the probes useful in further fingerprinting uses. For example, characteristic features of genetic sequences will be identified as being diagnostic of the entire sequence. However, in most embodiments, longer probe and target will be used, and for which slight mismatching may not need to be resolved.

**B. Preparation of Substrate Matrix**

[0116] A collection of specific probes may be produced by either of the methods described above in the section on sequencing. Specific oligonucleotide probes of desired lengths may be individually synthesized on a standard oligonucleotide synthesizer. The length of these probes is limited only by the length of the ability of the synthesizer to continue to accurately synthesize a molecule. Oligonucleotides or sequence fragments may also be isolated from natural sources. Biological amplification methods may be coupled with synthetic synthesizing procedures such as, e.g., polymerase chain reaction.

[0117] In one embodiment, the individually isolated probes may be attached to the matrix at defined positions. These probe reagents may be attached by an automated process using photochemical reagents, see, e.g., Dattagupta et al. (1985) U.S. Pat. No. 4,542,102 and (1987) U.S. Pat. No. 4,713,326. Each individual purified reagent can be attached individually at specific locations on a substrate.

[0118] In another embodiment, the VLSIPS synthesizing technique may be used to synthesize the desired probes at specific positions on a substrate. The probes may be synthesized by successively adding appropriate monomer subunits, e.g., nucleotides, to generate the desired sequences.

[0119] In another embodiment, a relatively short specific oligonucleotide is used which serves as a targeting reagent for positionally directing the sequence recognition reagent. For example, the sequence specific reagents having a separate additional sequence recognition segment (usually of a different polymer from the target sequence) can be directed to target oligonucleotides attached to the substrate. By use of non-natural targeting reagents, e.g., unusual nucleotide analogues which pair with other unnatural nucleotide analogues and which do not interfere with natural nucleotide interactions, the natural and non-natural portions can coexist on the same molecule without interfering with their individual functionalities. This can combine both a synthetic and biological production system analogous to the technique for targeting monoclonal antibodies to locations on a VLSIPS substrate at defined positions. Unnatural optical isomers of nucleotides may be useful unnatural reagents subject to similar chemistry, but incapable of interfering with the natural biological polymers.

[0120] After the separate substrate attached reagents are attached to the targeting segment, the two are crosslinked, thereby permanently attaching them to the substrate. Suitable crosslinking reagents are known, see, e.g., Dattagupta et al. (1985) U.S. Pat. No. 4,542,102 and (1987) "Coupling of nucleic acids to solid support by photochemical methods," U.S. Pat. No. 4,713,326. Similar linkages for attachment of proteins to a solid substrate are provided, e.g., in Merrifield (1986) *Science* 232:341.

**C. Labeling Target Nucleotides**

[0121] The labeling procedures used in the sequencing embodiments will also be applicable in the fingerprinting embodiments. However, since the fingerprinting embodiments often will involve relatively large target molecules and relatively short oligonucleotide probes, the amount of signal necessary to incorporate into the target sequence may be less critical than in the sequencing applications. For example, a relatively long target with a relatively small number of labels per molecule may be easily amplified or detected because of the relatively large target molecule size.

[0122] In various embodiments, it may be desired to cleave the target into smaller segments as in the sequencing embodiments. The labeling procedures and cleavage techniques described in the sequencing embodiments would usually also be applicable here.

**D. Hybridization Conditions**

[0123] The hybridization conditions used in fingerprinting embodiments will typically be less critical than for the sequencing embodiments. The reason is that the amount of mismatching which may be useful in providing the fingerprinting information would typically be far greater than that necessary in sequencing uses. For example, Southern hybridizations do not typically distinguish between slightly mismatched sequences. Under these circumstances, important and valuable information may be arrived at with less stringent hybridization conditions while providing valuable fingerprinting information. However, since the entire substrate is typically exposed to the target molecule at one time, the binding affinity of the probes should usually be of approximately comparable levels. For this reason, if oligonucleotide probes are being used, their lengths should be approximately comparable and will be selected to hybridize under conditions which are common for most of the probes on the substrate. Much as in a Southern hybridization, the target and oligonucleotide probes are of lengths typically greater than about 25 nucleotides. Under appropriate hybridization conditions, e.g., typically higher salt and lower temperature, the probes will hybridize irrespective of imperfect complementarity. In fact, with probes of greater than, e.g., about fifty nucleotides, the difference in stability of different sized probes will be relatively minor.

[0124] Typically the fingerprinting is merely for probing similarity or homology. Thus, the stringency of hybridization

can usually be decreased to fairly low levels. See, e.g., Wetmur and Davidson (1968) "Kinetics of Renaturation of DNA, " J. Mol. Biol., 31:349-370; and Kanehisa, M. (1984) Nuc. Acids Res., 12:203-213.

#### E. Detection; VLSIPS Scanning

[0125] Detection methods will be selected which are appropriate for the selected label. The scanning device need not necessarily be digitized or placed into a specific digital database, though such would most likely be done. For example, the analysis in fingerprinting could be photographic. Where a standardized fingerprint substrate matrix is used, the pattern of hybridizations may be spatially unique and may be compared photographically. In this manner, each sample may have a characteristic pattern of interactions and the likelihood of identical patterns will preferably be such low frequency that the fingerprint pattern indeed becomes a characteristic pattern virtually as unique as an individual's fingertip fingerprint. With a standardized substrate, every individual could be, in theory, uniquely identifiable on the basis of the pattern of hybridizing to the substrate.

[0126] Of course, the VLSIPS scanning apparatus may also be useful to generate a digitized version of the fingerprint pattern. In this way, the identification pattern can be provided in a linear string of digits. This sequence could also be used for a standardized identification system providing significant useful medical transferability of specific data. In one embodiment, the probes used are selected to be of sufficiently high resolution to measure polynucleotides encoding antigens of the major histocompatibility complex, it might even be possible to provide transplantation matching data in a linear stream of data. The fingerprinting data may provide a condensed version, or summary, of the linear genetic data, or any other information data base.

#### F. Analysis

[0127] The analysis of the fingerprint will often be much simpler than a total sequence determination. However, there may be particular types of analysis which will be substantially simplified by a selected group of probes. For example, probes which exhibit particular populational heterogeneity may be selected. In this way, analysis may be simplified and practical utility enhanced merely by careful selection of the specific probes and a careful matrix layout of those probes.

#### G. Substrate Reuse

[0128] As with the sequencing application, the fingerprinting usages may also take advantage of the reusability of the substrate. In this way, the interactions can be disrupted, the substrate treated, and the renewed substrate is equivalent to an unused substrate.

#### H. Other Polynucleotide Aspects

[0129] Besides using the fingerprinting method for analyzing the structure of a particular polynucleotide, the fingerprinting method may be used to characterize various samples. For example, a cell or population of cells may be tested for their expression of particular mRNA sequences, or for patterns of expressed mRNA species. This may be applicable to a cell or tissue type, to the expressed messenger RNA population expressed by a cell to the genetic content of a cell.

[0130] RNA can be isolated from a cell or a cell population, such as a purified cell fraction or a biopsy sample. The RNA may be labeled, for example by attaching a fluorescent molecule to isolated RNA or by using radiolabeled RNA (e.g., end-labeled with T4 polynucleotide kinase). A VLSIPS substrate containing positionally discrete oligonucleotide sequences may then be exposed to the pool of labeled RNA species under conditions permitting specific hybridization.

[0131] The pattern of positions at which labeled RNA has formed specific hybrids may be compared to a reference pattern to identify, and in some embodiments quantify, the expressed RNA species, or to identify the hybridization pattern itself as being characteristic of a particular cell type.

[0131] For example but not for limitation, a VLSIPS oligonucleotide substrate may be hybridized to a labeled RNA sample obtained from a first cell type (e.g., human lymphocytes) to establish a reference hybridization pattern for the first cell type. Similarly, an identical VLSIPS oligonucleotide substrate may be hybridized to a labeled RNA sample obtained from a second cell type (e.g., human monocytes) to establish a reference hybridization pattern for the second cell type. Labeled RNA may then be prepared from a cell or a cell population and hybridized to an identical VLSIPS oligonucleotide substrate, and the resultant hybridization pattern can be compared to the reference hybridization patterns established for the first and second cell types. By such comparisons, the RNA expression pattern of a cell or cell population can be identified as being similar to or distinct from one or more reference hybridization patterns.

[0132] Where a positionally discrete oligonucleotide on the VLSIPS substrate is in molar excess over the amount of the cognate (complementary) labeled RNA species in the hybridization reaction, the amount of specific hybridization to that VLSIPS locus (as measured by labeling intensity at that locus) can provide a quantitative measurement of the

cognate RNA species present in the labeled RNA sample. Thus, hybridization of labeled RNA to a VLSIPS oligonucleotide substrate can provide information identifying the individual RNA species that are expressed in a particular cell or cell population, as well as the relative abundance of one or more individual RNA species. This information can serve to fingerprint specific cell types or particular stages in cell differentiation.

5 [0133] For example but not for limitation, RNA samples prepared from tissue biopsies, specifically tumor biopsies, can be labeled and hybridized to a VLSIPS oligonucleotide substrate, and the resultant hybridization pattern can provide information regarding cell type, degree of differentiation, and metastatic potential (malignancy). Some of the positionally distinct oligonucleotides may hybridize specifically with RNA species transcribed from endogenous proto-oncogens (e.g., c-myc, c-ras<sup>H</sup>, c-sis, etc.) which are, in certain instances, transcribed at elevated levels in neoplastic tissues.

10 [0134] In addition to diagnostic applications, labeled RNA samples from various neoplastic cell types may be hybridized to VLSIPS oligonucleotide substrates and the resultant hybridization pattern(s) compared to reference patterns obtained with RNA from related, non-neoplastic cell types. Identification of distinctions between the hybridization patterns obtained with RNA from neoplastic cells as compared to patterns obtained from RNA from non-neoplastic cells may be of diagnostic value and may identify RNA species that encode proteins that are potential targets for novel

15 therapeutic modalities. In fact, the high resolution of the test will allow more complete characterization of parameters which define particular diseases. Thus, the power of diagnostic tests may be limited by the extent of statistical correlation with a particular condition rather than with the number of RNA species which are tested. The present invention provides the means to generate this large universe of possible reagents and the ability to actually accumulate that correlative data.

20 [0135] For fingerprinting of RNA expression patterns, the VLSIPS substrate polynucleotides will be at least 12 nucleotides in length, preferably at least 15 nucleotides in length, more preferably at least 25 nucleotides in length. The sequences of the positionally distinct polynucleotides on the VLSIPS substrate may be selected from published sources of sequence data, including but not limited to computerized database such as GenBank, and may or may not include random or pseudorandom sequences for detecting RNA species which have not yet been identified in the art. Fingerprint 25 analysis of RNA expression patterns will typically employ high-stringency washes so as to provide hybridization patterns that reflect predominantly specific hybridization. However, some nonspecific hybridization and/or cross-hybridization to slightly mismatched sequences may be tolerated, and in some embodiments may be desirable.

30 [0136] The ability to generate a high density means for screening the presence or absence of specific interactions allows for the possibility of screening for, if not saturating, all of a very large number of possible interactions. This is very powerful in providing the means for testing the combinations of molecular properties which can define a class of samples. For example, a species of organism may be characterized by its DNA sequences, e.g., a genetic fingerprint. By using a fingerprinting method, it may be determined that all members of that species are sufficiently similar in specific 35 sequences that they can be easily identified as being within a particular group. Thus, newly defined classes may be resolved by their similarity in fingerprint patterns. Alternatively, a non-member of that group will fail to share those many identifying characteristics. However, since the technology allows testing of a very large number of specific interactions, it also provides the ability to more finely distinguish between closely related different cells or samples. This will have important applications in diagnosing viral, bacterial, and other pathological or nonpathological infections.

40 [0137] In particular, cell classification may be defined by any of a number of different properties. For example, a cell class may be defined by its DNA sequences contained therein. This allows species identification for parasitic or other infections. For example, the human cell is presumably genetically distinguishable from a monkey cell, but different human cells will share many genetic markers. At higher resolution, each individual human genome will exhibit unique 45 sequences that can define it as a single individual.

[0138] Likewise, a developmental stage of a cell type may be definable by its pattern of expression of messenger RNA. For example, in particular stages of cells, high levels of ribosomal RNA are found whereas relatively low levels of other types of messenger RNAs may be found. The high resolution distinguishability provided by this fingerprinting method allows the distinction between cells which have relatively minor differences in its expressed mRNA population. Where a pattern is shown to be characteristic of a stage, a stage may be defined by that particular pattern of messenger RNA expression.

50 [0139] In another embodiment, a substrate as provided herein may be used for genetic screening. This would allow for simultaneous screening of thousands of genetic markers. As the density of the matrix is increased, many more molecules can be simultaneously tested. Genetic screening then becomes a simpler method as the present invention provides the ability to screen for thousands, tens of thousands, and hundreds of thousands, even millions of different possible genetic features. However, the number of high correlation genetic markers for conditions numbers only in the hundreds. Again, the possibility for screening a large number of sequences provides the opportunity for generating the 55 data which can provide correlation between sequences and specific conditions or susceptibility. The present invention provides the means to generate extremely valuable correlations useful for the genetic detection of the causative mutation leading to medical conditions. In still another embodiment, the present invention would be applicable to distinguishing two individuals having identical genetic compositions. The antibody population within an individual is depend-

ent both on genetic and historical factors. Each individual experiences a unique exposure to-various infectious agents, and the combined antibody expression is partly determined thereby. Thus, individuals may also be fingerprinted by their lymphocyte DNA or RNA hybridization pattern(s). Similar sorts of immunological and environmental histories may be useful for fingerprinting, perhaps in combination with other screening properties.

5 [0140] With the definition of new classes of cells, a cell sorter will be used to purify them. Moreover, new markers for defining that class of cells will be identified. For example, where the class is defined by its RNA content, cells may be screened by antisense probes which detect the presence or absence of specific sequences therein. Alternatively, cell lysates may provide information useful in correlating intracellular properties with extracellular markers which indicate functional differences. Using standard cell sorter technology with a fluorescence or labeled antisense probe which

10 recognizes the internal presence of the specific sequences of interest, the cell sorter will be able to isolate a relatively homogeneous population of cells possessing the particular marker. Using successive probes the sorting process should be able to select for cells having a combination of a large number of different markers.

15 [0141] With the fingerprinted method as in identification means arises from mosaicism problems in an organism. A mosaic organism is one whose genetic content in different cells is significantly different. Various clonal populations should have similar genetic fingerprints, though different clonal populations may have different genetic contents. See, for example, Suzuki et al. An Introduction to Genetic Analysis (4th Ed.), Freeman and Co., New York. However, this problem should be a relatively rare problem and could be more carefully evaluated with greater experience using the fingerprinting methods.

20 [0142] The invention will also find use in detecting changes, both genetic and in protein expression (i.e., by RNA expression fingerprinting), in a rapidly "evolving" protozoan infection, or similarly changing organism.

#### V. MAPPING

##### A. General

25 [0143] The use of the present invention for mapping parallels its use for fingerprinting and sequencing. Mapping provides the ability to locate particular segments along the length of the polynucleotide. The mapping provides the ability to locate, in a relative sense, the order of various subsequences. This may be achieved using at least two different approaches.

30 [0144] The first approach is to take the large sequence and fragment it at specific points. The fragments are then ordered and attached to a solid substrate. For example, the clones resulting from a chromosome walking process may be individually attached to the substrate by methods, e.g., caged biotin techniques, indicated earlier. Segments of unknown map position will be exposed to the substrate and will hybridize to the segment which contains that particular sequence. This procedure allows the rapid determination of a number of different labeled segments, each mapping requiring only a single hybridization step once the substrate is generated. The substrate may be regenerated by removal of the interaction, and the next mapping segment applied.

35 [0145] In an alternative method, a plurality of subsequences can be attached to a substrate. Various short probes may be applied to determine which segments may contain particular overlaps. The theoretical basis and a description of this mapping procedure is contained in, e.g., Evans et al. 1989 "Physical Mapping of Complex Genomes by Cosmid Multiplex Analysis," Proc. Natl. Acad. Sci. USA 86:5030-5034, and other references cited above in the Section labeled "Overall Description." Using this approach, the details of the mapping embodiment are very similar to those used in the fingerprinting embodiment.

##### B. Preparation of Substrate Matrix

45 [0146] The substrate may be generated in either of the methods generally applicable in the sequencing and fingerprinting embodiments. The substrate may be made either synthetically, or by attaching otherwise purified probes or sequences to the matrix. The probes or sequences may be derived either from synthetic or biological means. As indicated above, the solid phase substrate synthetic methods may be utilized to generate a matrix with positionally defined sequences. In the mapping embodiment, the importance of saturation of all possible subsequences of a pre-selected length is far less important than in the sequencing embodiment, but the length of the probes used may be desired to be much longer. The processes for making a substrate which has longer oligonucleotide probes should not be significantly different from those described for the sequencing embodiments, but the optimization parameters may be modified to comply with the mapping needs.

##### C. Labeling

55 [0147] The labeling methods will be similar to those applicable in sequencing and fingerprinting embodiments. Again,

the target sequences may be desired to be fragmented.

#### D. Hybridization/Specific Interaction

5 [0148] The specificity of interaction between the targets and probe would typically be closer to those used for fingerprinting embodiments, where homology is more important than absolute distinguishability of high fidelity complementary hybridization. Usually, the hybridization conditions will be such that merely homologous segments will interact and provide a positive signal. Much like the fingerprinting embodiment, it may be useful to measure the extent of homology by successive incubations at higher stringency conditions. Or, a plurality of different probes, each having various levels  
10 of homology may be used. In either way, the spectrum of homologies can be measured.

#### E. Detection

15 [0149] The detection methods used in the mapping procedure will be virtually identical to those used in the fingerprinting embodiment. The detection methods will be selected in combination with the labeling methods.

#### F. Analysis

20 [0150] The analysis of the data in a mapping embodiment will typically be somewhat different from that in fingerprinting. The fingerprinting embodiment will test for the presence or absence of specific or homologous segments. However, in the mapping embodiment, the existence of an interaction is coupled with some indication of the location of the interaction. The interaction is mapped in some manner to the physical polymer sequence. Some means for determining the relative positions of different probes is performed. This may be achieved by synthesis of the substrate in pattern, or may result from analysis of sequences after they have been attached to the substrate.

25 [0151] For example, the probes may be randomly positioned at various locations on the substrate. However, the relative positions of the various reagents in the original polymer may be determined by using short fragments, e.g., individually, as target molecules which determine the proximity of different probes. By an automated system of testing each different short fragment of the original polymer, coupled with proper analysis, it will be possible to determine which probes are adjacent one another on the original target sequence and correlate that with positions on the matrix. In this way, the matrix is useful for determining the relative locations of various new segments in the original target molecule. This sort of analysis is described in Evans, and the related references described above.

30 [0152] In another form of mapping, as described above in the fingerprinting section, the developmental map of a cell or biological system may be measured using fingerprinting type technology. Thus, the mapping may be along a temporal dimension rather than along a polymer dimension. The mapping or fingerprinting embodiments may also be used in determining the genetic rearrangements which may be genetically important, as in lymphocyte and B-cell development. In another example, various rearrangements or chromosomal dislocations may be tested by either the fingerprinting or mapping methods. These techniques are similar in many respects and the fingerprinting and mapping embodiments may overlap in many respects.

#### 40 G. Substrate Reuse

45 [0153] The substrate should be reusable in the manner described in the fingerprinting section. The substrate is renewed by removal of the specific interactions and is washed and prepared for successive cycles of exposure to new target sequences.

### VI. ADDITIONAL SCREENING AND APPLICATIONS

#### A. Specific Interactions

50 [0154] The production of a high density plurality of spatially segregated polymers provides the ability to generate a very large universe or repertoire of individually and distinct sequence possibilities. As indicated above, particular oligonucleotides may be synthesized in automated fashion at specific locations on a matrix. In fact, these oligonucleotides may be used to direct other molecules to specific locations by linking specific oligonucleotides to other reagents which are in batch exposed to the matrix and hybridized in a complementary fashion to only those locations where the complementary oligonucleotide has been synthesized on the matrix. This allows for spatially attaching a plurality of different reagents onto the matrix instead of individually attaching each separate reagent at each specific location. Although the caged biotin method allows the automated attachment, the speed of the caged biotin attachment process is relatively slow and requires a separate reaction for each reagent being attached. By use of the oligonucleotide method, the

specificity of position can be done in an automated and parallel fashion. As each reagent is produced, instead of directly attaching each reagent at each desired position, the reagent may be attached to a specific desired complementary oligonucleotide which will ultimately be specifically directed toward locations on the matrix having a complementary oligonucleotide attached thereto.

[0155] In addition, the technology allows screening for specificity of interaction with particular reagents. For example, the oligonucleotide sequence specificity of binding of a potential reagent may be tested by presenting to the reagent all of the possible subsequences available for binding. Although secondary or higher order sequence specific features might not be easily screenable using this technology, it does provide a convenient, simple, quick, and thorough screen of interactions between a reagent and its target recognition sequences. See, e.g., Pfeifer et al. (1989) Science 246: 810-812.

[0156] For example, the interaction of a promoter protein with its target binding sequence may be tested for many different, or all, possible binding sequences. By testing the strength of interactions under various different conditions, the interaction of the promoter protein with each of the different potential binding sites may be analyzed. The spectrum of strength of interactions with each different potential binding site may provide significant insight into the types of features which are important in determining specificity.

[0157] An additional example of a sequence specific interaction between reagents is the testing of binding of a double stranded nucleic acid structure with a single stranded oligonucleotide. Often, a triple stranded structure is produced which has significant aspects of sequence specificity. Testing of such interactions with either sequences comprising only natural nucleotides, or perhaps the testing of nucleotide analogs may be very important in screening for particularly useful diagnostic or therapeutic reagents. See, e.g., Häner and Dervan (1990) Biochemistry 29:9761-6765, and references therein.

#### B. Sequence Comparisons

[0158] Once a gene is sequenced, the present invention provides means to compare alleles or related sequences to locate and identify differences from the control sequence. This would be extremely useful in further analysis of genetic variability at a specific gene locus.

#### C. Categorizations

[0159] As indicated above in the fingerprinting and mapping embodiments, the present invention is also useful to define specific stages in the temporal sequence of cells, e.g., development, and the resulting tissues within an organism. For example, the developmental stage of a cell, or population of cells, can be dependent upon the expression of particular messenger RNAs. The screening procedures provided allow for high resolution definition of new classes of cells. In addition, the temporal development of particular cells will be characterized by the presence or expression of various mRNAs. Means to simultaneously screen a plurality or very large number of different sequences as provided. The combination of different markers made available dramatically increases the ability to distinguish fairly closely related cell types. Other markers may be combined with markers and methods made available herein to define new classifications of biological samples, e.g., based upon new combinations of markers.

[0160] The presence or absence of particular marker sequences will be used to define temporal developmental stages. Once the stages are defined, fairly simple methods can be applied to actually purify those particular cells. For example, antisense probes or recognition reagents may be used with a cell sorter to select those cells containing or expressing the critical markers. Alternatively, the expression of those sequences may result in specific antigens which may also be used in defining cell classes and sorting those cells away from others. In this way, for example, it should be possible to select a class of omnipotent immune system cells which are able to completely regenerate a human immune system. Based upon the cellular classes defined by the parameters made available by this technology, purified classes of cells having identifiable differences in RNA expression and/or DNA structure are made available.

[0161] In an alternative embodiment, subclasses of T-cells are defined, in part, upon the combination of expressed cell surface RNA species. The present invention allows for the simultaneous screening of a large plurality of different RNA species together. Thus, higher resolution classification of different T-cell subclasses becomes possible and, with the definitions and functional differences which correlate with those other parameters, the ability to purify those cell types becomes available. This is applicable not only to T-cells, lymphocyte cells, or even to freely circulating cells. Many of the cells for which this would be most useful will be immobile cells found in particular tissues or organs. Tumor cells will be diagnosed or detected using these fingerprinting techniques. Coupled with a temporal change in structure, developmental classes may also be selected and defined using these technologies. The present invention also provides the ability not only to define new classes of cells based upon functional or structural differences, but it also provides the ability to select or purify populations of cells which share these particular properties. In particular, antisense DNA or RNA molecules may be introduced into a cell to detect RNA sequences therein. See, e.g., Weintraub (1990) Scientific

American 262:40-46.

D. Statistical Correlations

5 [0162] In an additional embodiment, the present invention also allows for the high resolution correlation of medical conditions with various different markers. For example, the present technology, when applied to amniocentesis or other genetic screening methods, typically screen for tens of different markers at most. The present invention allows simultaneous screening for tens, hundreds, thousands, tens of thousands, hundreds of thousands, and even millions of different genetic sequences. Thus, applying the fingerprinting methods of the present invention to a sufficiently large population allows detailed statistical analysis to be made, thereby correlating particular medical conditions with particular markers, typically genetic markers or pathognomonic RNA expression patterns. Tumor-specific RNA expression patterns and particular RNA species characterizing various neoplastic phenotypes will be identified using the present invention.

10 [0163] Various medical conditions may be correlated against an enormous data base of the sequences within an individual. Genetic propensities and correlations then become available and high resolution genetic predictability and correlation become much more easily performed. With the enormous data base, the reliability of the predictions also is better tested. Particular markers which are partially diagnostic of particular medical conditions or medical susceptibilities will be identified and provide direction in further studies and more careful analysis of the markers involved. Of course, as indicated above in the sequencing embodiment, the present invention will find much use in intense sequencing projects. For example, sequencing of the entire human genome in the human genome project will be greatly simplified and enabled by the present invention.

15

20

**VI. FORMATION OF SUBSTRATE**

25 [0164] The substrate is provided with a pattern of specific reagents which are positionally localized on the surface of the substrate. This matrix of positions is defined by the automated system which produces the substrate. The instrument will typically be one similar to that described in PCT publication no. WO90/15070. The instrumentation described therein is directly applicable to the applications used here. In particular, the apparatus comprises a substrate, typically a silicon containing substrate, on which positions on the surface may be defined by a coordinate system of positions. These positions can be individually addressed or detected by the VLSIPS apparatus.

30 [0165] Typically, the VLSIPS apparatus uses optical methods used in semiconductor fabrication applications. In this way, masks may be used to photo-activate positions for attachment or synthesis of specific sequences on the substrate. These manipulations may be automated by the types of apparatus described in PCT publication no. WO90/15070.

35 [0166] Selectively removable protecting groups allow creation of well defined areas of substrate surface having differing reactivities. Preferably, the protecting groups are selectively removed from the surface by applying a specific activator, such as electromagnetic radiation of a specific wavelength and intensity. More preferably, the specific activator exposes selected areas of surface to remove the protecting groups in the exposed areas.

40 [0167] Protecting groups of the present invention are used in conjunction with solid phase oligonucleotide syntheses using deoxyribonucleic and ribonucleic acids. In addition to protecting the substrate surface from unwanted reaction, the protecting groups block a reactive end of the monomer to prevent self-polymerization.

45 [0168] Attachment of a protecting group to the 5'-hydroxyl group of a nucleoside during synthesis using for example, phosphate-triester coupling chemistry, prevents the 5'-hydroxyl of one nucleoside from reacting with the 3'-activated phosphate-triester of another.

50 [0169] Regardless of the specific use, protecting groups are employed to protect a moiety on a molecule from reacting with another reagent. Protecting groups of the present invention have the following characteristics: they prevent selected reagents from modifying the group to which they are attached; they are stable (that is, they remain attached) to the synthesis reaction conditions; they are removable under conditions that do not adversely affect the remaining structure; and once removed, do not react appreciably with the surface or surface-bound oligonucleotide.

55 [0170] In a preferred embodiment, the protecting groups will be photoactivatable. The properties and uses of photoreactive protecting compounds have been reviewed. See, McCray et al., *Ann. Rev. of Biophys. and Biophys. Chem.* (1989) 18:239-270. Preferably, the photosensitive protecting groups will be removable by radiation in the ultraviolet (UV) or visible portion of the electromagnetic spectrum. More preferably, the protecting groups will be removable by radiation in the near UV or visible portion of the spectrum. In some embodiments, however, activation may be performed by other methods such as localized heating, electron beam lithography, laser pumping, oxidation or reduction with microelectrodes, and the like. Sulfonyl compounds are suitable reactive groups for electron beam lithography. Oxidative or reductive removal is accomplished by exposure of the protecting group to an electric current source, preferably using microelectrodes directed to the predefined regions of the surface which are desired for activation.

60 [0171] The density of reagents attached to a silicon substrate may be varied by standard procedures. The surface

area for attachment of reagents may be increased by modifying the silicon surface. For example, a matte surface may be machined or etched on the substrate to provide more sites for attachment of the particular reagents. Another way to increase the density of reagent binding sites is to increase the derivitization density of the silicon. Standard procedures for achieving this are described, below.

5 [0172] One method to control the derivatization density is to highly derivatize the substrate with photochemical groups at high density. The substrate is then photolyzed for various predetermined times, which photoactivate the groups at a measurable rate, and react then with a capping reagent. By this method, the density of linker groups may be modulated by using a desired time and intensity of photoactivation.

10 [0173] In many applications, the number of different sequences which may be provided may be limited by the density and the size of the substrate on which the matrix pattern is generated. In situations where the density is insufficiently high to allow the screening of the desired number of sequences, multiple substrates may be used to increase the number of sequences tested. Thus, the number of sequences tested may be increased by using a plurality of different substrates. Because the VLSIPS apparatus is almost fully automated, increasing the number of substrates does not lead to a significant increase in the number of manipulations which must be performed by humans. This again leads to greater reproducibility and speed in the handling of these multiple substrates.

#### A. Instrumentation

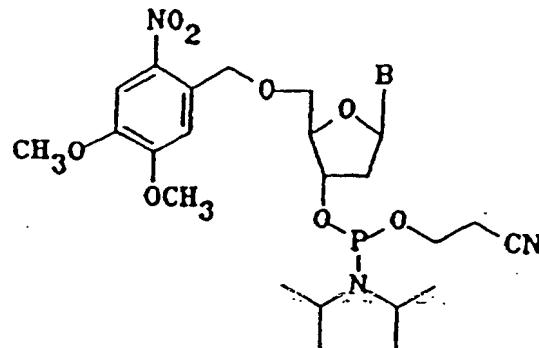
20 [0174] The concept of using VLSIPS generally allows a pattern or a matrix of reagents to be generated. The procedure for making the pattern is performed by any of a number of different methods. An apparatus and instrumentation useful for generating a high density VLSIPS substrate is described in detail in PCT publication no. WO90/15070.

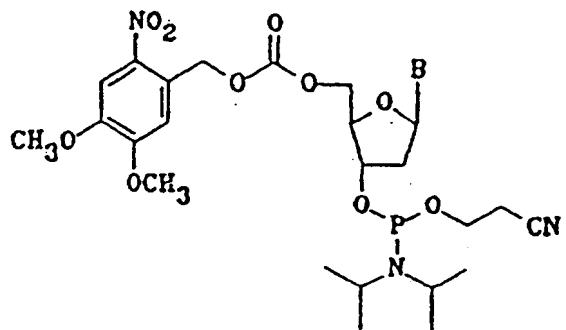
#### B. Binary Masking

25 [0175] For example, the binary masking technique allows for producing a plurality of sequences based on the selection of either of two possibilities at any particular location. By a series of binary masking steps, the binary decision may be the determination, on a particular synthetic cycle, whether or not to add any particular one of the possible subunits. By treating various regions of the matrix pattern in parallel, the binary masking strategy provides the ability to carry out spatially addressable parallel synthesis.

#### C. Synthetic Methods

30 [0176] The construction of the matrix pattern on the substrate will typically be generated by the use of photo-sensitive reagents. By use of photo-lithographic optical methods, particular segments of the substrate can be irradiated with light to activate or deactivate blocking agents, e.g., to protect or deprotect particular chemical groups. By an appropriate sequence of photo-exposure steps at appropriate times with appropriate masks and with appropriate reagents, the substrates can have known polymers synthesized at positionally defined regions on the substrate. Methods for synthesizing various substrates are described in PCT publication no. WO90/15070. By a sequential series of these photo-exposure and reaction manipulations, a defined matrix pattern of known sequences may be generated, and is typically referred to as a VLSIPS substrate. In the nucleic acid synthesis embodiment, nucleosides used in the synthesis of DNA by photolytic methods will typically be one of the two forms shown below:



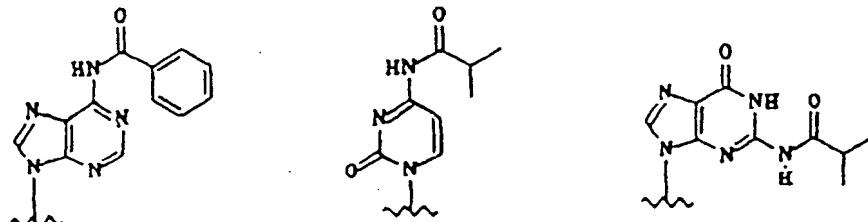


[0177] B = Adenine, Cytosine, Guanine, or Thymine

20 [0178] In I, the photolabile group at the 5' position is abbreviated NV (nitroveratryl) and in II, the group is abbreviated NVOC (nitroveratryl oxycarbonyl). Although not shown above, bases (adenine, cytosine, and guanine) contain exocyclic NH<sub>2</sub> groups which must be protected during DNA synthesis. Thymine contains no exocyclic NH<sub>2</sub> and therefore requires no protection. The standard protecting groups for these amines are shown below:

25

30



35

40 Adenine (A)

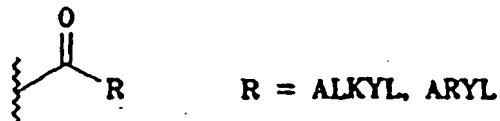
Cytosine (C)

Guanine (G)

45

[0179] Other amides of the general formula

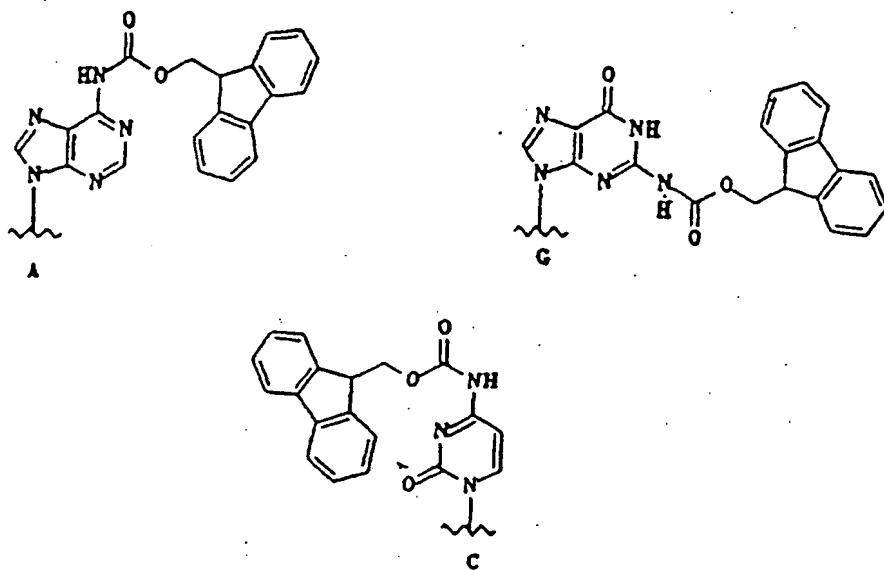
50



55

where R may be alkyl or aryl have been used.

[0180] Another type of protecting group Fmoc (9-fluorenyl methoxycarbonyl) is currently being used to protect the exocyclic amines of the three bases:



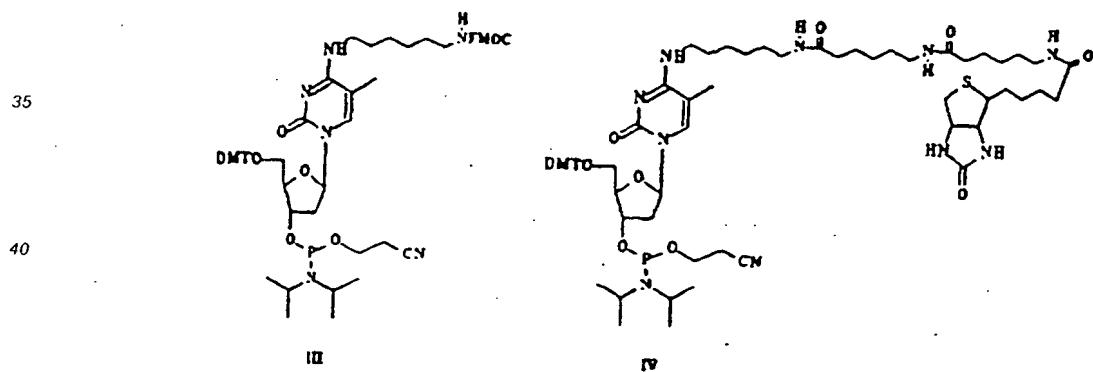
Adenine (A)

Cytosine (C)

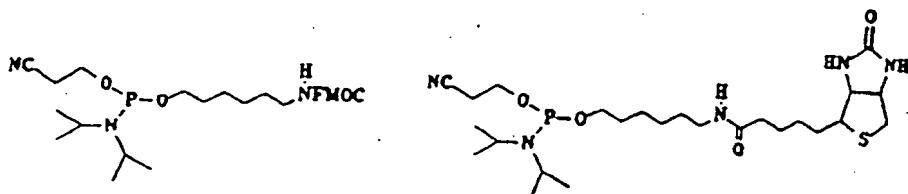
Guanine (G)

25 [0181] The advantage of the Fmoc group is that it is removed under mild conditions (dilute organic bases) and can be used for all three bases. The amide protecting groups require more harsh conditions to be removed (NH<sub>3</sub>/MeOH with heat).

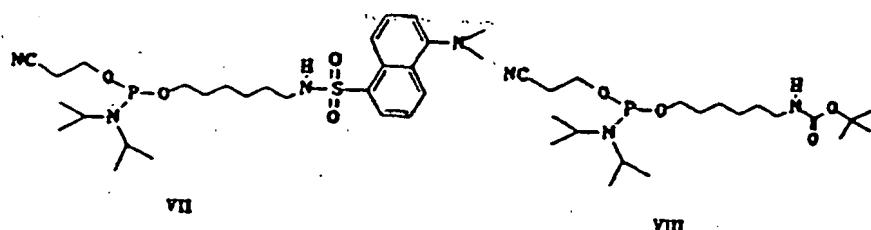
30 [0182] Nucleosides used as 5'-OH probes, useful in verifying correct VLSIPS synthetic function, have been the following:



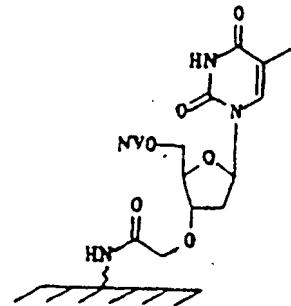
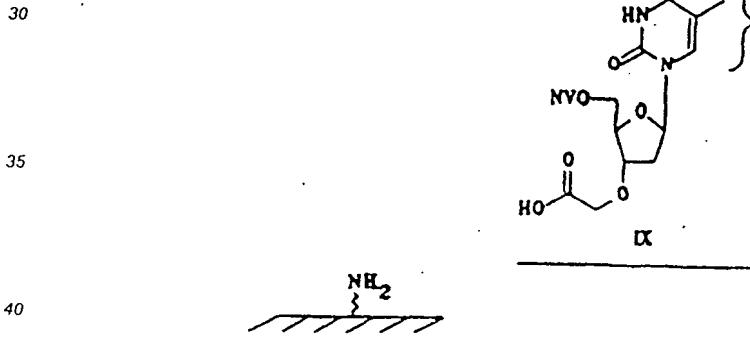
45 [0183] These compounds are used to detect where on a substrate photolysis has occurred by the attachment of either III or V to the newly generated 5'-OH. In the case of III, after the phosphate attachment is made, the substrate is treated with a dilute base to remove the Fmoc group. The resulting amine can be reacted with FITC and the substrate examined by fluorescence microscopy. This indicates the proper generation of a 5'-OH. In the case of compound IV, after the phosphate attachment is made, the substrate is treated with FITC labeled streptavidin and the substrate again may be examined by fluorescence microscopy. Other probes, although not nucleoside based, have included the following:



10



[0184] The method of attachment of the first nucleoside to the surface of the substrate depends on the functionality of the groups at the substrate surface. If the surface is amine functionalized, an amide bond is made (see example below).



[0185] If the surface is hydroxy functionalized a phosphate bond is made (see example below)

45

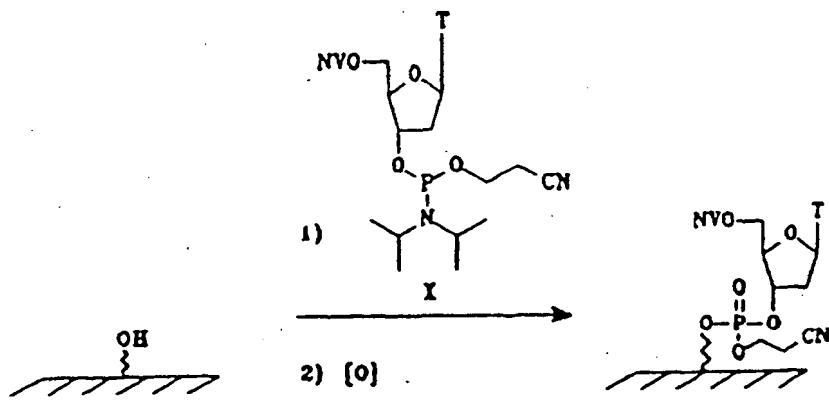
50

55

5

10

15



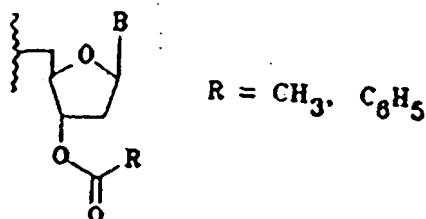
[0186] In both cases, the thymidine example is illustrated, but any one of the four phosphoramidite activated nucleosides can be used in the first step.

[0187] Photolysis of the photolabile group NV or NVOC on the 5' positions of the nucleosides is carried out at -362 nm with an intensity of 14 mW/cm<sup>2</sup> for 10 minutes with the substrate side (side containing the photolabile group) immersed in dioxane. After the coupling of the next nucleoside is complete, the photolysis is repeated followed by another coupling until the desired oligomer is obtained.

[0188] One of the most common 3'-O-protecting group is the ester, in particular the acetate.

25

30



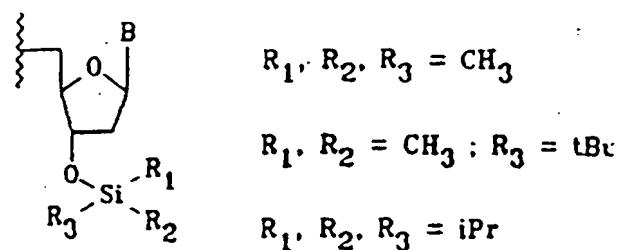
35

[0189] The groups can be removed by mild base treatment 0.1N NaOH/MeOH or K<sub>2</sub>CO<sub>3</sub>/H<sub>2</sub>O/MeOH.

[0190] Another group used most often is the silyl ether.

40

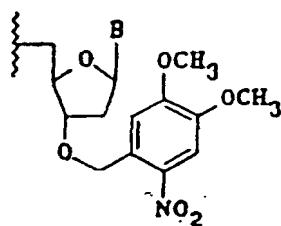
45



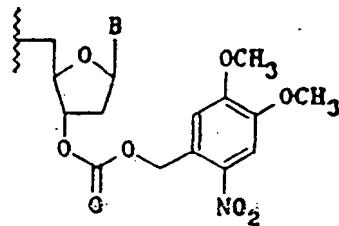
50

[0191] These groups can be removed by neutral conditions using 1 M tetra-n-butylammonium fluoride in THF or under acid conditions.

[0192] Related to photodeprotection, the nitroveratryl group could also be used to protect the 3'-position.



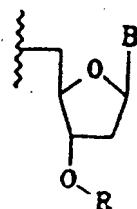
OR



10

[0193] Here, light (photolysis) would be used to remove these protecting groups.  
 [0194] A variety of ethers can also be used in the protection of the 3'-O-position.

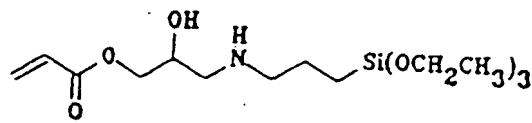
15

**R = TRITYL, BENZYL**

25

[0195] Removal of these groups usually involves acid or catalytic methods.  
 [0196] Although the specificity of interactions at particular locations will usually be homogeneous due to a homogeneous polymer being synthesized at each defined location, for certain purposes, it may be useful to have mixed polymers with a commensurate mixed collection of interactions occurring at specific defined locations, or degeneracy reducing analogues, which have been discussed above and show broad specificity in binding. Then, a positive interaction signal may result from any of a number of sequences contained therein.  
 [0197] As an alternative method of generating a matrix pattern on a substrate, preformed polymers may be individually attached at particular sites on the substrate. This may be performed by individually attaching reagents one at a time to specific positions on the matrix, a process which may be automated. Another way of generating a positionally defined matrix pattern on a substrate is to have individually specific reagents which interact with each specific position on the substrate. For example, oligonucleotides may be synthesized at defined locations on the substrate. Then the substrate would have on its surface a plurality of regions having homogeneous oligonucleotides attached at each position.  
 [0198] In particular, at least four different substrate preparation procedures are available for treating a substrate surface. They are the standard VLSIPS method, polymeric substrates, Durapore™, and synthetic beads or fibers. The treatment labeled "standard VLSIPS" method involves applying aminopropyltrietoxysilane to a glass surface.  
 [0199] The polymeric substrate approach involves either of two ways of generating a polymeric substrate. The first uses a high concentration of aminopropyltrietoxysilane (2-20%) in an aqueous ethanol solution (95%). This allows the silane compound to polymerize both in solution and on the substrate surface, which provides a high density of amines on the surface of the glass. This density is contrasted with the standard VLSIPS method. This polymeric method allows for the deposition on the substrate surface of a monolayer due to the anhydrous method used with the aforementioned silane.  
 [0200] The second polymeric method involves either the coating or covalent binding of an appropriate acrylic acid polymer onto the substrate surface. In particular, e.g., in DNA synthesis, a monomer such as a hydroxylpropylacrylate is used to generate a high density of hydroxyl groups on the substrate surface, allowing for the formation of phosphate bonds. An example of such a compound is shown:

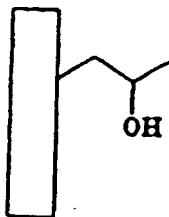
55



[0201] The method using a Durapore™ membrane (Millipore) consists of a polyvinylidene difluoride coating with crosslinked polyhydroxylpropyl acrylate [PVDF-HPA]:

5

10



15 Here the building up of, e.g., a DNA oligomer, can be started immediately since phosphate bonds to the surface can be accomplished in the first step with no need for modification. A nucleotide dimer (5'-C-T-3') has been successfully made on this substrate in our labs.

[0202] The fourth method utilizes synthetic beads or fibers. This would use another substrate, such as a teflon copolymer graft bead or fiber, which is covalently coated with an organic layer (hydrophilic) terminating in hydroxyl sites (commercially available from Molecular Brosystems, Inc.) This would offer the same advantage as the Durapore™ membrane, allowing for immediate phosphate linkages, but would give additional contour by the 3-dimensional growth of oligomers.

[0203] A matrix pattern of new reagents may be targeted to each specific oligonucleotide position by attaching a complementary oligonucleotide to which the substrate bound form is complementary. For instance, a number of regions may have homogeneous oligonucleotides synthesized at various locations. Oligonucleotide sequences complementary to each of these can be individually generated and linked to a particular specific reagents. Often these specific reagents will be antibodies. As each of these is specific for finding its complementary oligonucleotide, each of the specific reagents will bind through the oligonucleotide to the appropriate matrix position. A single step having a combination of different specific reagents being attached specifically to a particular oligonucleotide will thereby bind to its complement at the defined matrix position. The oligonucleotides will typically then be covalently attached, using, e.g., an acridine dye, for photocrosslinking. Psoralen is a commonly used acridine dye for photocrosslinking purposes, see, e.g., Song et al. (1979) *Photochem. Photobiol.* 29:1177-1197; Cimino et al. (1985) *Ann. Rev. Biochem.* 54:1151-1193; Parsons (1980) *Photochem. Photobiol.* 32:813-821; and Dattagupta et al. (1985) U.S. Pat. No. 4,542,102, and (1987) U.S. Pat. No. 4,713,326. This method allows a single attachment manipulation to attach all of the specific reagents to the matrix at defined positions and results in the specific reagents being homogeneously located at defined positions.

35

#### D. Surface Immobilization

##### 1. caged biotin

[0204] An alternative method of attaching reagents in a positionally defined matrix pattern is to use a caged biotin system. In short, the caged biotin has a photosensitive blocking moiety which prevents the combination of avidin to biotin. At positions where the photo-lithographic process has removed the blocking group, high affinity biotin sites are generated. Thus, by a sequential series of photolithographic deblocking steps interspersed with exposure of those regions to appropriate biotin containing reagents, only those locations where the deblocking takes place will form an avidin-biotin interaction. Because the avidin-biotin binding is very tight, this will usually be virtually irreversible binding.

##### 2. crosslinked interactions

[0205] The surface immobilization may also take place by photocrosslinking of defined oligonucleotides linked to specific reagents. After hybridization of the complementary oligonucleotides, the oligonucleotides may be crosslinked by a reagent by psoralen or another similar type of acridine dye. Other useful crosslinking reagents are described in Dattagupta et al. (1985) U.S. Pat. No. 4,542,102, and (1987) U.S. Pat. No. 4,713,326.

[0206] In another embodiment, colony or phage plaque transfer of biological polymers may be transferred directly onto a silicon substrate. For example, a colony plate may be transferred onto a substrate having a generic oligonucleotide sequence which hybridizes to another generic complementary sequence contained on all of the vectors into which inserts are cloned. This will specifically only bind those molecules which are actually contained in the vectors containing the desired complementary sequence. This immobilization allows for producing a matrix onto which a sequence specific reagent can bind, or for other purposes. In a further embodiment, a plurality of different vectors each having a specific

oligonucleotide attached to the vector may be specifically attached to particular regions on a matrix having a complementary oligonucleotide attached thereto.

5           VIII. HYBRIDIZATION/SPECIFIC INTERACTION

A. General

[0207] As discussed previously in the VLSIPS parent applications, the VLSIPS substrates may be used for screening for specific interactions with sequence specific targets or probes.

[0208] In addition, the availability of substrates having the entire repertoire of possible sequences of a defined length opens up the possibility of sequencing by hybridization. This sequence may be de novo determination of an unknown sequence, particularly of nucleic acid, verification of a sequence determined by another method, or an investigation of changes in a previously sequenced gene, locating and identifying specific changes. For example, often Maxam and Gilbert sequencing techniques are applied to sequences which have been determined by Sanger and Coulson. Each of those sequencing technologies have problems with resolving particular types of sequences. Sequencing by hybridization may serve as a third and independent method for verifying other sequencing techniques. See, e.g., (1988) *Science* 242:1245.

[0209] In addition, the ability to provide a large repertoire of particular sequences allows use of short subsequence and hybridization as a means to fingerprint a polynucleotide sample. For example, fingerprinting to a high degree of specificity of sequence matching may be used for identifying highly similar samples, e.g., those exhibiting high homology to the selected probes. This may provide a means for determining classifications of particular sequences. This should allow determination of whether particular genomes of bacteria, phage, or even higher cells might be related to one another.

[0210] In addition, fingerprinting may be used to identify an individual source of biological sample. See, e.g., Lander, E. (1989) *Nature*, 339:501-505, and references therein. For example, a DNA fingerprint may be used to determine whether a genetic sample arose from another individual. This would be particularly useful in various sorts of forensic tests to determine, e.g., paternity or sources of blood samples. Significant detail on the particulars of genetic fingerprinting for identification purposes are described in, e.g., Morris et al. (1989) "Biostatistical evolution of evidence from continuous allele frequency distribution DNA probes in reference to disputed paternity of identity," *J. Forensic Science* 34:1311-1317; and Neufeld et al. (1990) *Scientific American* 262:46-53.

[0211] In another embodiment, a fingerprinting-like procedure may be used for classifying cell types by analyzing a pattern of specific nucleic acids present in the cell, specifically RNA expression patterns. This may also be useful in defining the temporal stage of development of cells, e.g., stem cells or other cells which undergo temporal changes in development. For example, the stage of a cell, or group of cells, may be tested or defined by isolating a sample of mRNA from the population and testing to see what sequences are present in messenger populations. Direct samples, or amplified samples (e.g., by polymerase chain reaction), may be used. Where particular mRNA or other nucleic acid sequences may be characteristic of or shown to be characteristic of particular developmental stages, physiological states, or other conditions, this fingerprinting method may define them.

[0212] The present invention may also be used for mapping sequences within a larger segment. This may be performed by at least two methods, particularly in reference to nucleic acids. Often, enormous segments of DNA are subcloned into a large plurality of subsequences. Ordering these subsequences may be important in determining the overlaps of sequences upon nucleotide determinations. Mapping may be performed by immobilizing particularly large segments onto a matrix using the VLSIPS technology. Alternatively, sequences may be ordered by virtue of subsequences shared by overlapping segments. See, e.g., Craig et al. (1990) *Nuc. Acids Res.* 18:2653-2660; Michiels et al. (1987) *CABIOS* 3:203-210; and Olson et al. (1986) *Proc. Natl. Acad. Sci. USA* 83:7826-7830.

B. Important Parameters

[0213] The extent of specific interaction between reagents immobilized to the VLSIPS substrate and another sequence specific reagent may be modified by the conditions of the interaction. Sequencing embodiments typically require high fidelity hybridization and the ability to discriminate perfect matching from imperfect matching. Fingerprinting and mapping embodiments may be performed using less stringent conditions, or in some embodiments very highly stringent conditions, depending upon the circumstances.

[0214] In a nucleic acid hybridization embodiment, the specificity and kinetics of hybridization have been described in detail by, e.g., Welmur and Davidson (1968) *J. Mol. Biol.*, 31:349-370, Britten and Kohne (1968) *Science* 161: 529-530, and Kanehisa, (1984) *Nuc. Acids Res.* 12:203-213. Parameters which are well known to affect specificity and kinetics of reaction include salt conditions, ionic composition of the solvent, hybridization temperature, length of oligonucleotide matching sequences, guanine and cytosine (GC) content, presence of hybridization accelerators, pH, spe-

cific bases found in the matching sequences, solvent conditions, and addition of organic solvents.

[0215] In particular, the salt conditions required for driving highly mismatched sequences to completion typically include a high salt concentration. The typical salt used is sodium chloride (NaCl), however, other ionic salts may be utilized, e.g., KCl. Depending on the desired stringency hybridization, the salt concentration will often be less than about 3 molar, more often less than 2.5 molar, usually less than about 2 molar, and more usually less than about 1.5 molar. For applications directed towards higher stringency matching, the salt concentrations would typically be lower. Ordinary high stringency conditions will utilize salt concentration of less than about 1 molar, more often less than about 750 millimolar, usually less than about 500 millimolar, and may be as low as about 250 or 150 millimolar.

[0216] The kinetics of hybridization and the stringency of hybridization both depend upon the temperature at which the hybridization is performed and the temperature at which the washing steps are performed. Temperatures at which steps for low stringency hybridization are desired would typically be lower temperatures, e.g., ordinarily at least about 15°C, more ordinarily at least about 20°C, usually at least about 25°C, and more usually at least about 30°C. For those applications requiring high stringency hybridization, or fidelity of hybridization and sequence matching, temperatures at which hybridization and washing steps are performed would typically be high. For example, temperatures in excess of about 35°C would often be used, more often in excess of about 40°C, usually at least about 45°C, and occasionally even temperatures as high as about 50°C or 60°C or more. Of course, the hybridization of oligonucleotides may be disrupted by even higher temperatures. Thus, for stripping of targets from substrates, as discussed below, temperatures as high as 80°C, or even higher may be used.

[0217] The base composition of the specific oligonucleotides involved in hybridization affects the temperature of melting, and the stability of hybridization as discussed in the above references. However, the bias of GC rich sequences to hybridize faster and retain stability at higher temperatures can be compensated for by the inclusion in the hybridization incubation or wash steps of various buffers. Sample buffers which accomplish this result include the triethyl- and trimethyl ammonium buffers. See, e.g., Wood et al. (1987) *Proc. Natl. Acad. Sci. USA*, 82:1585-1588, and Khrapko, K. et al. (1989) *FEBS Letters* 256:118-122.

[0218] The rate of hybridization can also be affected by the inclusion of particular hybridization accelerators. These hybridization accelerators include the volume exclusion agents characterized by dextran sulfate, or polyethylene glycol (PEG). Dextran sulfate is typically included at a concentration of between 1% and 40% by weight. The actual concentration selected depends upon the application, but typically a faster hybridization is desired in which the concentration is optimized for the system in question. Dextran sulfate is often included at a concentration of between 0.5% and 2% by weight or dextran sulfate at a concentration between about 0.5% and 5%. Alternatively, proteins which accelerate hybridization may be added, e.g., the recA protein found in *E. coli* or other homologous proteins.

[0219] Of course, the specific hybridization conditions will be selected to correspond to a discriminatory condition which provides a positive signal where desired but fails to show a positive signal at affinities where interaction is not desired. This may be determined by a number of titration steps or with a number of controls which will be run during the hybridization and/or washing steps to determine at what point the hybridization conditions have reached the stage of desired specificity.

## IX. DETECTION METHODS

[0220] Methods for detection depend upon the label selected. The criteria for selecting an appropriate label are discussed below, however, a fluorescent label is preferred because of its extreme sensitivity and simplicity. Standard labeling procedures are used to determine the positions where interactions between a sequence and a reagent take place. For example, if a target sequence is labeled and exposed to a matrix of different probes, only those locations where probes do interact with the target will exhibit any signal. Alternatively, other methods may be used to scan the matrix to determine where interaction takes place. Of course, the spectrum of interactions may be determined in a temporal manner by repeated scans of interactions which occur at each of a multiplicity of conditions. However, instead of testing each individual interaction separately, a multiplicity of sequence interactions may be simultaneously determined on a matrix.

### A. Labeling Techniques

[0221] The target polynucleotide may be labeled by any of a number of convenient detectable markers. A fluorescent label is preferred because it provides a very strong signal with low background. It is also optically detectable at high resolution and sensitivity through a quick scanning procedure. Other potential labeling moieties include, radioisotopes, chemiluminescent compounds, labeled binding proteins, heavy metal atoms, spectroscopic markers, magnetic labels, and linked enzymes.

[0222] Another method for labeling does not require incorporation of a labeling moiety. The target may be exposed to the probes, and a double strand hybrid is formed at those positions only. Addition of a double strand specific reagent

will detect where hybridization takes place. An intercalative dye such as ethidium bromide may be used as long as the probes themselves do not fold back on themselves to a significant extent forming hairpin loops. See, e.g., Sheldon et al. (1986) U.S. Pat. No. 4,582,789. However, the length of the hairpin loops in short oligonucleotide probes would typically be insufficient to form a stable duplex.

5 [0223] In another embodiment, different targets may be simultaneously sequenced where each target has a different label. For instance, one target could have a green fluorescent label and a second target could have a red fluorescent label. The scanning step will distinguish sites of binding of the red label from those binding the green fluorescent label. Each sequence can be analyzed independently from one another.

10 [0224] Suitable chromogens will include molecules and compounds which absorb light in a distinctive range of wavelengths so that a color may be observed, or emit light when irradiated with radiation of a particular wave length or wave length range, e.g., fluorescers.

15 [0225] A wide variety of suitable dyes are available, being primary chosen to provide an intense color with minimal absorption by their surroundings. Illustrative dye types include quinoline dyes, triarylmethane dyes, acridine dyes, alizarine dyes, phthaleins, insect dyes, azo dyes, anthraquinoid dyes, cyanine dyes, phenazathionium dyes, and phen-azoxonium dyes.

20 [0226] A wide variety of fluorescers may be employed either by themselves or in conjunction with quencher molecules. Fluorescers of interest fall into a variety of categories having certain primary functionalities. These primary functionalities include 1- and 2-aminonaphthalene, p,p'-diaminostilbenes, pyrenes, quaternary phenanthridine salts, 9-aminoacridines, p,p'-diaminobenzopnenone imines, anthracenes, oxacarbocyanine, merocyanine, 3-aminoequilen-25 in, perylene, bis-benzoxazole, bis-p-oxazolyl benzene, 1,2-benzophenazin, retinol, bis-3-aminopyridinium salts, helle-brigenin, tetracycline, sterophenol, benzimidzaolylphenylamine, 2-oxo-3-chromen, indole, xanthen, 7-hydroxycoumarin, phenoxazine, salicylate, strophanthidin, porphyrins, triarylmethanes and flavin. Individual fluorescent compounds which have functionalities for linking or which can be modified to incorporate such functionalities include, e.g., dansyl chloride; fluoresceins such as 3,6-dihydroxy-9-phenylxanthhydrol; rhodamineisothiocyanate; N-phenyl 1-amino-8-sulfonatonaphthalene; N-phenyl 2-amino-6-sulfonatonaphthalene; 4-acetamido-4-isothiocyanato-stilbene-2,2'-disulfonic acid; pyrene-3-sulfonic acid; 2-toluidinonaphthalene-6-sulfonate; N-phenyl, N-methyl 2-aminoaphthalene-6-sulfonate; ethidium bromide; stebrine; auromine-0,2-(9'-anthroyl)palmitate; dansyl phosphatidylethanolamine; N,N'-diocadecyl oxacarbocyanine; N,N'-dihexyl oxacarbocyanine; merocyanine 4-(3'pyrenyl)butyrate; d-3-aminodesoxy-equilenin; 12-(9'-anthroyl)stearate; 2-methylanthracene; 9-vinylanthracene; 2,2'-(vinylene-p-phenylene)bisbenzoxazole; p-bis[2-(4-methyl-5-phenyl-oxazolyl)]benzene; 6-dimethylamino-1,2-benzophenazin; retinol; bis(3'-aminopyridinium) 1,10-decandiyli diiodide; sulfonaphthylhydrazone of hellbrienin; chlorotetracycline; N-(7-dimethylamino-4-methyl-2-oxo-3-chromenyl)maleimide; N-[p-(2-benzimidazolyl)-phenyl]maleimide; N-(4-fluoranthyl)maleimide; bis(homovanillic acid); resazarin; 4-chloro-7-nitro-2,1,3-benzooxadiazole; merocyanine 540; resorufin; rose bengal; and 2,4-diphenyl-3(2H)-furanone.

35 [0227] Desirably, fluorescers should absorb light above about 300 nm, preferably about 350 nm, and more preferably above about 400 nm, usually emitting at wavelengths greater than about 10 nm higher than the wavelength of the light absorbed. It should be noted that the absorption and emission characteristics of the bound dye may differ from the unbound dye. Therefore, when referring to the various wavelength ranges and characteristics of the dyes, it is intended to indicate the dyes as employed and not the dye which is unconjugated and characterized in an arbitrary solvent.

40 [0228] Fluorescers are generally preferred because by irradiating a fluorescer with light, one can obtain a plurality of emissions. Thus, a single label can provide for a plurality of measurable events.

45 [0229] Detectable signal may also be provided by chemiluminescent and bioluminescent sources. Chemiluminescent sources include a compound which becomes electronically excited by a chemical reaction and may then emit light which serves as the detectable signal or donates energy to a fluorescent acceptor. A diverse number of families of compounds have been found to provide chemiluminescence under a variety of conditions. One family of compounds is 2,3-dihydro-1,-4-phthalazinedione. The most popular compound is luminol, which is the 5-amino compound. Other members of the family include the 5-amino-6,7,8-trimethoxy- and the dimethylamino[ca]benz analog. These compounds can be made to luminesce with alkaline hydrogen peroxide or calcium hypochlorite and base. Another family of compounds is the 2,4,5-triphenylimidazoles, with lophine as the common name for the parent product. Chemiluminescent analogs include para-dimethylamino and -methoxy substituents. Chemiluminescence may also be obtained with oxalates, usually oxalyl active esters, e.g., p-nitrophenyl and a peroxide, e.g., hydrogen peroxide, under basic conditions. Alternatively, luciferins may be used in conjunction with luciferase or lucigenins to provide bioluminescence.

50 [0230] Spin labels are provided by reporter molecules with an unpaired electron spin which can be detected by electron spin resonance (ESR) spectroscopy. Exemplary spin labels include organic free radicals, transitional metal complexes, particularly vanadium, copper, iron, and manganese, and the like. Exemplary spin labels include nitroxide free radicals.

B. Scanning System

[0231] With the automated detection apparatus, the correlation of specific positional labeling is converted to the presence on the target of sequences for which the reagents have specificity of interaction. Thus, the positional information is directly converted to a database indicating what sequence interactions have occurred. For example, in a nucleic acid hybridization application, the sequences which have interacted between the substrate matrix and the target molecule can be directly listed from the positional information. The detection system used is described in PCT publication no. WO90/15070. Although the detection described therein is a fluorescence detector, the detector may be replaced by a spectroscopic or other detector. The scanning system may make use of a moving detector relative to a fixed substrate, a fixed detector with a moving substrate, or a combination. Alternatively, mirrors or other apparatus can be used to transfer the signal directly to the detector.

[0232] The detection method will typically also incorporate some signal processing to determine whether the signal at a particular matrix position is a true positive or may be a spurious signal. For example, a signal from a region which has actual positive signal may tend to spread over and provide a positive signal in an adjacent region which actually should not have one. This may occur, e.g., where the scanning system is not properly discriminating with sufficiently high resolution in its pixel density to separate the two regions. Thus, the signal over the spatial region may be evaluated pixel by pixel to determine the locations and the actual extent of positive signal. A true positive signal should, in theory, show a uniform signal at each pixel location. Thus, processing by plotting number of pixels with actual signal intensity should have a clearly uniform signal intensity. Regions where the signal intensities show a fairly wide dispersion, may be particularly suspect and the scanning system may be programmed to more carefully scan those positions.

[0233] In another embodiment, as the sequence of a target is determined at a particular location, the overlap for the sequence would necessarily have a known sequence. Thus, the system can compare the possibilities for the next adjacent position and look at these in comparison with each other. Typically, only one of the possible adjacent sequences should give a positive signal and the system might be programmed to compare each of these possibilities and select that one which gives a strong positive. In this way, the system can also simultaneously provide some means of measuring the reliability of the determination by indicating what the average signal to background ratio actually is.

[0234] More sophisticated signal processing techniques can be applied to the initial determination of whether a positive signal exists or not.

[0235] From a listing of those sequences which interact, data analysis may be performed on a series of sequences. For example, in a nucleic acid sequence application, each of the sequences may be analyzed for their overlap regions and the original target sequence may be reconstructed from the collection of specific subsequences obtained therein. Other sorts of analyses for different applications may also be performed, and because the scanning system directly interfaces with a computer the information need not be transferred manually. This provides for the ability to handle large amounts of data with very little human intervention. This, of course, provides significant advantages over manual manipulations. Increased throughput and reproducibility is thereby provided by the automation of vast majority of steps in any of these applications.

## DATA ANALYSIS

A. General

[0236] Data analysis will typically involve aligning the proper sequences with their overlaps to determine the target sequence. Although the target "sequence" may not specifically correspond to any specific molecule, especially where the target sequence is broken and fragmented up in the sequencing process, the sequence corresponds to a contiguous sequence of the subfragments.

[0237] The data analysis can be performed by a computer using an appropriate program. See, e.g., Drmanac, R. et al. (1989) *Genomics* 4:114-128; and a commercially available analysis program available from the Genetic Engineering Center, P.O. Box 794, 11000 Belgrade, Yugoslavia. Although the specific manipulations necessary to reassemble the target sequence from fragments may take many forms, one embodiment uses a sorting program to sort all of the subsequences using a defined hierarchy. The hierarchy need not necessarily correspond to any physical hierarchy, but provides a means to determine, in order, which subfragments have actually been found in the target sequence. In this manner, overlaps can be checked and found directly rather than having to search throughout the entire set after each selection process. For example, where the oligonucleotide probes are 10-mers, the first 9 positions can be sorted. A particular subsequence can be selected as in the examples, to determine where the process starts. As analogous to the theoretical example provided above, the sorting procedure provides the ability to immediately find the position of the subsequence which contains the first 9 positions and can compare whether there exists more than 1 subsequence during the first 9 positions. In fact, the computer can easily generate all of the possible target sequences which contain given combination of subsequences. Typically there will be only one, but in various situations, there will be more.

5 [0238] An exemplary flow chart for a sequencing program is provided in Figure 1. In general terms, the program provides for automated scanning of the substrate to determine the positions of probe and target interaction. Simple processing of the intensity of the signal may be incorporated to filter out clearly spurious signals. The positions with positive interaction are correlated with the sequence specificity of specific matrix positions, to generate the set of matching subsequences. This information is further correlated with other target sequence information, e.g., restriction fragment analysis. The sequences are then aligned using overlap data, thereby leading to possible corresponding target sequences which will, optimally, correspond to a single target sequence.

10 **B. Hardware**

15 [0239] A variety of computer systems may be used to run a sequencing program. The program may be written to provide both the detecting and scanning steps together and will typically be dedicated to a particular scanning apparatus. However, the components and functional steps may be separated and the scanning system may provide an output, e.g., through tape or an electronic connection into a separate computer which separately runs the sequencing analysis program. The computer may be any of a number of machines provided by standard computer manufacturers, e.g., IBM compatible machines, Apple™ machines, VAX machines, and others, which may often use a UNIX™ operating system. Alternatively, custom computing architectures may be employed, these architectures may include neural network methods implemented in hardware and/or software. Of course, the hardware used to run the analysis program will typically determine what programming language would be used.

20 **C. Software**

25 [0240] Software would be readily developed by a person of ordinary skill in the programming art, following the flow chart provided, or based upon the input provided and the desired result.

[0241] Of course, an exemplary embodiment is a polynucleotide sequence system. However, the theoretical and mathematical manipulations necessary for data analysis of other linear molecules are conceptually similar.

XI. SUBSTRATE REUSE

30 [0242] Where a substrate is made with specific reagents that are relatively insensitive to the handling and processing steps involved in a single cycle of use, the substrate may often be reused. The target molecules are usually stripped off of the solid phase specific recognition molecules. Of course, it is preferred that the manipulations and conditions be selected as to be mild and to not affect the substrate. For example, if a substrate is acid labile, a neutral pH would be preferred in all handling steps. Similar sensitivities would be carefully respected where recycling is desired.

35 **A. Removal of Label**

40 [0243] Typically for a recycling, the previously attached specific interaction would be disrupted and removed. This will typically involve exposing the substrate to conditions under which the interaction between probe and target is disrupted. Alternatively, it may be exposed to conditions where the target is destroyed. For example, where the probes are oligonucleotides and the target is a polynucleotide, a heating and low salt wash will often be sufficient to disrupt the interactions. Additional reagents may be added such as detergents, and organic or inorganic solvents which disrupt the interaction between the specific reagents and target.

45 **B. Storage and Preservation**

50 [0244] As indicated above, the matrix will typically be maintained under conditions where the matrix itself and the linkages and specific reagents are preserved. Various specific preservatives may be added which prevent degradation. For example, if the reagents are acid or base labile, a neutral pH buffer will typically be added. It is also desired to avoid destruction of the matrix by growth of organisms which may destroy organic reagents attached thereto. For this reason, a preservative such as cyanide or azide may be added. However, the chemical preservative should also be selected to preserve the chemical nature of the linkages and other components of the substrate. Typically, a detergent may also be included.

55 **C. Processes to Avoid Degradation of Oligomers**

[0245] In particular, a substrate comprising a large number of oligomers will be treated in a fashion which is known to maintain the quality and integrity of oligonucleotides. These include storing the substrate in a carefully controlled

environment under conditions of lower temperature, cation depletion (EDTA and EGTA), sterile conditions, and inert argon or nitrogen atmosphere.

## XII. INTEGRATED SEQUENCING STRATEGY

5

### A. Initial Mapping Strategy

[0246] As indicated above, although the VLSIPS may be applied to sequencing embodiments, it is often useful to integrate other concepts to simplify the sequencing. For example, nucleic acids may be easily sequenced by careful selection of the vectors and hosts used for amplifying and generating the specific target sequences. For example, it may be desired to use specific vectors which have been designed to interact most efficiently with the VLSIPS substrate. This is also important in fingerprinting and mapping strategies. For example, vectors may be carefully selected having particular complementary sequences which are designed to attach to a genetic or specific oligomer on the substrate. This is also applicable to situations where it is desired to target particular sequences to specific locations on the matrix.

[0247] In one embodiment, unnatural oligomers may be used to target natural probes to specific locations on the VLSIPS substrate. In addition, particular probes may be generated for the mapping embodiment which are designed to have specific combinations of characteristics. For example, the construction of a mapping substrate may depend upon use of another automated apparatus which takes clones isolated from a chromosome walk and attaches them individually or in bulk to the VLSIPS substrate.

[0248] In another embodiment, a variety of specific vectors having known and particular "targeting" sequences adjacent the cloning sites may be individually used to clone a selected probe, and the isolated probe will then be targetable to a site on the VLSIPS substrate with a sequence complementary to the "target" sequence.

25

### B. Selection of Smaller Clones

[0249] In the fingerprinting and mapping embodiments, the selection of probes may be very important. Significant mathematical analysis may be applied to determine which specific sequences should be used as those probes. Of course, for fingerprinting use, sequences that show significant heterogeneity across the human population would be preferred. Selection of the specific sequences which would most favorably be utilized will tend to be single copy sequences within the genome, and more specifically single copy sequences that have low cross-hybridization potential to other sequences in the genome (i.e., not members of a closely-related multigene family).

[0250] Various hybridization selection procedures may be applied to select sequences which tend not to be repeated within a genome, and thus would tend to be conserved across individuals. For example, hybridization selections may be made for non-repetitive and single copy sequences. See, e.g., Britten and Kohne (1968) "Repeated Sequences in DNA," *Science* 161:529-540. On the other hand, it may be desired under certain circumstances to use repeated sequences. For example, where a fingerprint may be used to identify or distinguish different species, or where repetitive sequences may be diagnostic of specific species, repetitive sequences may be desired for inclusion in the fingerprinting probes. In either case, the sequencing capability will greatly assist in the selection of appropriate sequences to be used as probes.

[0251] Also as indicated above, various means for constructing an appropriate substrate may involve either mechanical or automated procedures. The standard VLSIPS automated procedure involves synthesizing oligonucleotides or short polymers directly on the substrate. In various other embodiments, it is possible to attach separately synthesized reagents onto the matrix in an ordered array. Other circumstances may lend themselves to transfer a pattern from a petri plate onto a solid substrate. Also, there are methods for site specifically directing collections of reagents to specific locations using unnatural nucleotides or equivalent sorts of targeting molecules.

[0252] While a brute force manual transfer process may be utilized sequentially attaching various samples to successive positions, instrumentation for automating such procedures may also be devised. The automated system for performing such would preferably be relatively easily designed and conceptually easily understood.

50

## XIII. COMMERCIAL APPLICATIONS

55

### A. Sequencing

[0253] As indicated above, sequencing may be performed either *de novo* or as a verification of another sequencing method. The present hybridization technology provides the ability to sequence nucleic acids and polynucleotides *de novo*, or as a means to verify either the Maxam and Gilbert chemical sequencing technique or Sanger and Coulson dideoxy- sequencing techniques. The hybridization method is useful to verify sequencing determined by any other sequencing technique and to closely compare two similar sequences, e.g.; to identify and locate sequence differences.

[0254] Of course, sequencing of can be very important in many different sorts of environments. For example, it will be useful in determining the genetic sequence of particular markers in various individuals. In addition, polymers may be used as markers or for information containing molecules to encode information. For example, a short polynucleotide sequence may be included in large bulk production samples indicating the manufacturer, date, and location of manufacture of a product. For example, various drugs may be encoded with this information with a small number of molecules in a batch. For example, a pill may have somewhere from 10 to 100 to 1,000 or more very short and small molecules encoding this information. When necessary, this information may be decoded from a sample of the material using a polymerase chain reaction (PCR) or other amplification method. This encoding system may be used to provide the origin of large bulky samples without significantly affecting the properties of those samples. For example, chemical samples may also be encoded by this method thereby providing means for identifying the source and manufacturing details of lots. The origin of bulk hydrocarbon samples may be encoded. Production lots of organic compounds such as benzene or plastics may be encoded with a short molecule polymer. Food stuffs may also be encoded using similar marking molecules. Even toxic waste samples can be encoded determining the source or origin. In this way, proper disposal can be traced or more easily enforced.

[0255] Similar sorts of encoding may be provided by fingerprinting-type analysis. Whether the resolution is absolute or less so, the concept of coding information on molecules such as nucleic acids, which can be amplified and later decoded, may be a very useful and important application.

[0256] This technology also provides the ability to include markers for origins of biological materials. For example, a patented animal line may be transformed with a particular unnatural sequence which can be traced back to its origin. With a selection of multiple markers, the likelihood could be negligible that a combination of markers would have independently arisen from a source other than the patented or specifically protected source. This technique may provide a means for tracing the actual origin of particular biological materials. Bacteria, plants, and animals will be subject to marking by such encoding sequences.

## 25 B. Fingerprinting

[0257] As indicated above, fingerprinting technology may also be used for data encryption. Moreover, fingerprinting allows for significant identification of particular individuals. Where the fingerprinting technology is standardized, and used for identification of large numbers of people, related equipment and peripheral processing will be developed to accompany the underlying technology. For example, specific equipment may be developed for automatically taking a biological sample and generating or amplifying the information molecules within the sample to be used in fingerprinting analysis. Moreover, the fingerprinting substrate may be mass produced using particular types of automatic equipment. Synthetic equipment may produce the entire matrix simultaneously by stepwise synthetic methods as provided by the VLSIPS technology. The attachment of specific probes onto a substrate may also be automated.

[0258] In addition, peripheral processing may be important and may be dedicated to this specific application. Thus, automated equipment for producing the substrates may be designed, or particular systems which take in a biological sample and output either a computer readout or an encoded instrument, e.g., a card or document which indicates the information and can provide that information to others. An identification having a short magnetic strip with a few million bits may be used to provide individual identification and important medical information useful in a medical emergency.

[0259] In fact, data banks may be set up to correlate all of this information of fingerprinting with medical information. This may allow for the determination of correlations between various medical problems and specific DNA sequences. By collating large populations of medical records with genetic information, genetic propensities and genetic susceptibilities to particular medical conditions may be developed. Moreover, with standardization of substrates, the micro encoding data may be also standardized to reproduce the information from a centralized data bank or on an encoding device carried on an individual person. On the other hand, if the fingerprinting procedure is sufficiently quick and routine, every hospital may routinely perform a fingerprinting operation and from that determine many important medical parameters for an individual.

[0260] In particular industries, the VLSIPS sequencing, fingerprinting, or mapping technology will be particularly appropriate. As mentioned above, agricultural livestock suppliers may be able to encode and determine whether their particular strains are being used by others. By incorporating particular markers into their genetic stocks, the markers will indicate origin of genetic material. This is applicable to seed producers, livestock producers, and other suppliers of medical or agricultural biological materials.

[0261] This may also be useful in identifying individual animals or plants. For example, these markers may be useful in determining whether certain fish return to their original breeding grounds, whether sea turtles always return to their original birthplaces, or to determine the migration patterns and viability of populations of particular endangered species. It would also provide means for tracking the sources of particular animal products. For example, it might be useful for determining the origins of controlled animal substances such as elephant ivory or particular bird populations whose importation or exportation is controlled.

[0262] As indicated above, polymers may be used to encode important information on source and batch and supplier. This is described in greater detail, e.g., "Applications of PCR to industrial problems," (1990) in Chemical and Engineering News 68:145. In fact, the synthetic method can be applied to the storage of enormous amounts of information. Small substrates may encode enormous amounts of information, and its recovery will make use of the inherent replication capacity. For example, on regions of  $10\text{ }\mu\text{m} \times 10\text{ }\mu\text{m}$ ,  $1\text{ cm}^2$  has  $10^6$  regions. In theory, the entire human genome could be attached in 1000 nucleotide segments on a  $3\text{ cm}^2$  surface. Genomes of endangered species may be stored on these substrates.

[0263] Fingerprinting may also be used for genetic tracing or for identifying individuals for forensic science purposes. See, e.g., Morris, J. et al. (1989) "Biostatistical Evaluation of Evidence From Continuous Allele Frequency Distribution DNA Probes in Reference to Disputed Paternity and Identity," J. Forensic Science 34:1311-1317, and references provided therein.

[0264] In addition, the high resolution fingerprinting allows the distinguishability to high resolution of particular samples. As indicated above, new cell classifications may be defined based on combinations of a large number of properties. Similar applications will be found in distinguishing different species of animals or plants. In fact, microbial identification may become dependent or characterization of the genetic content. Tumors or other cells exhibiting abnormal physiology will be detectable by use of the present invention. Also, knowing the genetic fingerprint of a microorganism may provide very useful information on how to treat an infection by such organism.

[0265] Modifications of the fingerprint embodiments may be used to diagnose the condition of the organism. For example, a blood sample is presently used for diagnosing any of a number of different physiological conditions. A multi-dimensional fingerprinting method made available by the present invention could become a routine means for diagnosing an enormous number of physiological features simultaneously. This may revolutionize the practice of medicine in providing information on an enormous number of parameters together at one time. In another way, the genetic predisposition may also revolutionize the practice of medicine providing a physician with the ability to predict the likelihood of particular medical conditions arising at any particular moment. It also provides the ability to apply preventative medicine.

[0266] Also available are kits with the reagents useful for performing sequencing, fingerprinting, and mapping procedures. The kits will have various compartments with the desired necessary reagents, e.g., substrate, labeling reagents for target samples, buffers, and other useful accompanying products.

### 30 C. Mapping

[0267] The present invention also provides the means for mapping sequences within enormous stretches of sequence. For example, nucleotide sequences may be mapped within enormous chromosome size sequence maps. For example, it would be possible to map a chromosomal location within the chromosome which contains hundreds of millions of nucleotide base pairs. In addition, the mapping and fingerprinting embodiments allow for testing of chromosomal translocations, one of the standard problems for which amniocentesis is performed.

[0268] The present invention will be better understood by reference to the following illustrative examples. The following examples are offered by way of illustration and not by way of limitation.

[0269] Relevant techniques are described in PCT publication no. WO90/15070, published December 13, 1990; PCT publication no. WO91/07087, published May 30, 1991.

[0270] Also, additional relevant techniques are described, e.g., in Sambrook, J., et al. (1989) Molecular Cloning: a Laboratory Manual, 2d Ed., vols 1-3, Cold Spring Harbor Press, New York; Greenstein and Winitz (1961) Chemistry of the Amino Acids, Wiley and Sons, New York; Bodzansky, M. (1988) Peptide Chemistry: a Practical Textbook, Springer-Verlag, New York; Harlow and Lane (1988) Antibodies: A Laboratory Manual, Cold Spring Harbor Press, New York; Glover, D. (ed.) (1987) DNA Cloning: A Practical Approach, vols 1-3, IRL Press, Oxford; Bishop and Rawlings (1987) Nucleic Acid and Protein Sequence Analysis: A Practical Approach, IRL Press, Oxford; Hames and Higgins (1985) Nucleic Acid Hybridisation: A Practical Approach, IRL Press, Oxford; Wu et al. (1989) Recombinant DNA Methodology, Academic Press, San Diego; Goding (1986) Monoclonal Antibodies: Principles and Practice, (2d ed.), Academic Press, San Diego; Finegold and Barron (1986) Bailey and Scott's Diagnostic Microbiology, (7th ed.), Mosby Co., St. Louis; Collins et al. (1989) Microbiological Methods, (6th ed.), Butterworth, London; Chaplin and Kennedy (1986) Carbohydrate Analysis: A Practical Approach, IRL Press, Oxford; Van Dyke (ed.) (1985) Bioluminescence and Chemiluminescence: Instruments and Applications, vol 1, CRC Press, Boca Raton; and Ausubel, et al. (ed.) (1990) Current Protocols in Molecular Biology, Greene Publishing and Wiley-Interscience, New York.

### 55 EXAMPLES

[0271] The following examples are provided to illustrate the efficacy of the inventions herein. All operations were conducted at about ambient temperatures and pressures unless indicated to the contrary.

## POLYNUCLEOTIDE SEQUENCING

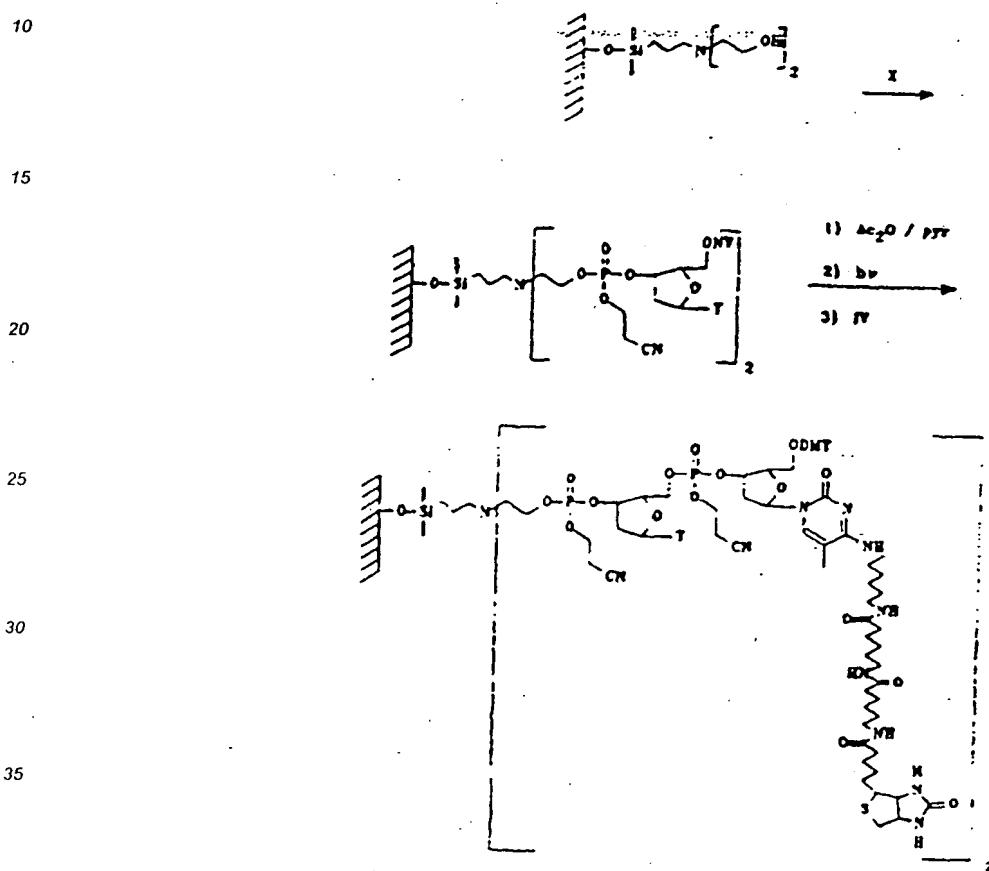
## 1. HPLC of the photolysis of 5'-O-nitroveratryl-thymidine.

5 [0272] In order to determine the time for photolysis of 5'-O-nitroveratryl thymidine a 100  $\mu$ M solution of NV-Thym-OH (5'-O-nitroveratryl thymidine) in dioxane was made and -200  $\mu$ l aliquots were irradiated (in a quartz cuvette 1 cm x 2 mm) at 362.3 nm for 20 sec, 40 sec, 60 sec, 2 min, 5 min, 10 min, 15 min, and 20 min. The resulting irradiated mixtures were then analyzed by HPLC using a Varian MicroPak SP column (C<sub>18</sub> analytical) at a flow rate of 1 ml/min and a solvent system of 40% CH<sub>3</sub>CN and 60% water. Thymidine has a retention time of 1.2 min and NVO-Thym-OH has a retention time of 2.1 min. It was seen that after 10 min of exposure the deprotection was complete.

## 10 15 20 25 30 35 40 45 50 55 107 112 117 122 127 132 137 142 147 152 157 162 167 172 177 182 187 192 197 202 207 212 217 222 227 232 237 242 247 252 257 262 267 272 277 282 287 292 297 302 307 312 317 322 327 332 337 342 347 352 357 362 367 372 377 382 387 392 397 402 407 412 417 422 427 432 437 442 447 452 457 462 467 472 477 482 487 492 497 502 507 512 517 522 527 532 537 542 547 552 557 562 567 572 577 582 587 592 597 602 607 612 617 622 627 632 637 642 647 652 657 662 667 672 677 682 687 692 697 702 707 712 717 722 727 732 737 742 747 752 757 762 767 772 777 782 787 792 797 802 807 812 817 822 827 832 837 842 847 852 857 862 867 872 877 882 887 892 897 902 907 912 917 922 927 932 937 942 947 952 957 962 967 972 977 982 987 992 997 1002 1007 1012 1017 1022 1027 1032 1037 1042 1047 1052 1057 1062 1067 1072 1077 1082 1087 1092 1097 1102 1107 1112 1117 1122 1127 1132 1137 1142 1147 1152 1157 1162 1167 1172 1177 1182 1187 1192 1197 1202 1207 1212 1217 1222 1227 1232 1237 1242 1247 1252 1257 1262 1267 1272 1277 1282 1287 1292 1297 1302 1307 1312 1317 1322 1327 1332 1337 1342 1347 1352 1357 1362 1367 1372 1377 1382 1387 1392 1397 1402 1407 1412 1417 1422 1427 1432 1437 1442 1447 1452 1457 1462 1467 1472 1477 1482 1487 1492 1497 1502 1507 1512 1517 1522 1527 1532 1537 1542 1547 1552 1557 1562 1567 1572 1577 1582 1587 1592 1597 1602 1607 1612 1617 1622 1627 1632 1637 1642 1647 1652 1657 1662 1667 1672 1677 1682 1687 1692 1697 1702 1707 1712 1717 1722 1727 1732 1737 1742 1747 1752 1757 1762 1767 1772 1777 1782 1787 1792 1797 1802 1807 1812 1817 1822 1827 1832 1837 1842 1847 1852 1857 1862 1867 1872 1877 1882 1887 1892 1897 1902 1907 1912 1917 1922 1927 1932 1937 1942 1947 1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 2002 2007 2012 2017 2022 2027 2032 2037 2042 2047 2052 2057 2062 2067 2072 2077 2082 2087 2092 2097 2102 2107 2112 2117 2122 2127 2132 2137 2142 2147 2152 2157 2162 2167 2172 2177 2182 2187 2192 2197 2202 2207 2212 2217 2222 2227 2232 2237 2242 2247 2252 2257 2262 2267 2272 2277 2282 2287 2292 2297 2302 2307 2312 2317 2322 2327 2332 2337 2342 2347 2352 2357 2362 2367 2372 2377 2382 2387 2392 2397 2402 2407 2412 2417 2422 2427 2432 2437 2442 2447 2452 2457 2462 2467 2472 2477 2482 2487 2492 2497 2502 2507 2512 2517 2522 2527 2532 2537 2542 2547 2552 2557 2562 2567 2572 2577 2582 2587 2592 2597 2602 2607 2612 2617 2622 2627 2632 2637 2642 2647 2652 2657 2662 2667 2672 2677 2682 2687 2692 2697 2702 2707 2712 2717 2722 2727 2732 2737 2742 2747 2752 2757 2762 2767 2772 2777 2782 2787 2792 2797 2802 2807 2812 2817 2822 2827 2832 2837 2842 2847 2852 2857 2862 2867 2872 2877 2882 2887 2892 2897 2902 2907 2912 2917 2922 2927 2932 2937 2942 2947 2952 2957 2962 2967 2972 2977 2982 2987 2992 2997 3002 3007 3012 3017 3022 3027 3032 3037 3042 3047 3052 3057 3062 3067 3072 3077 3082 3087 3092 3097 3102 3107 3112 3117 3122 3127 3132 3137 3142 3147 3152 3157 3162 3167 3172 3177 3182 3187 3192 3197 3202 3207 3212 3217 3222 3227 3232 3237 3242 3247 3252 3257 3262 3267 3272 3277 3282 3287 3292 3297 3302 3307 3312 3317 3322 3327 3332 3337 3342 3347 3352 3357 3362 3367 3372 3377 3382 3387 3392 3397 3402 3407 3412 3417 3422 3427 3432 3437 3442 3447 3452 3457 3462 3467 3472 3477 3482 3487 3492 3497 3502 3507 3512 3517 3522 3527 3532 3537 3542 3547 3552 3557 3562 3567 3572 3577 3582 3587 3592 3597 3602 3607 3612 3617 3622 3627 3632 3637 3642 3647 3652 3657 3662 3667 3672 3677 3682 3687 3692 3697 3702 3707 3712 3717 3722 3727 3732 3737 3742 3747 3752 3757 3762 3767 3772 3777 3782 3787 3792 3797 3802 3807 3812 3817 3822 3827 3832 3837 3842 3847 3852 3857 3862 3867 3872 3877 3882 3887 3892 3897 3902 3907 3912 3917 3922 3927 3932 3937 3942 3947 3952 3957 3962 3967 3972 3977 3982 3987 3992 3997 4002 4007 4012 4017 4022 4027 4032 4037 4042 4047 4052 4057 4062 4067 4072 4077 4082 4087 4092 4097 4102 4107 4112 4117 4122 4127 4132 4137 4142 4147 4152 4157 4162 4167 4172 4177 4182 4187 4192 4197 4202 4207 4212 4217 4222 4227 4232 4237 4242 4247 4252 4257 4262 4267 4272 4277 4282 4287 4292 4297 4302 4307 4312 4317 4322 4327 4332 4337 4342 4347 4352 4357 4362 4367 4372 4377 4382 4387 4392 4397 4402 4407 4412 4417 4422 4427 4432 4437 4442 4447 4452 4457 4462 4467 4472 4477 4482 4487 4492 4497 4502 4507 4512 4517 4522 4527 4532 4537 4542 4547 4552 4557 4562 4567 4572 4577 4582 4587 4592 4597 4602 4607 4612 4617 4622 4627 4632 4637 4642 4647 4652 4657 4662 4667 4672 4677 4682 4687 4692 4697 4702 4707 4712 4717 4722 4727 4732 4737 4742 4747 4752 4757 4762 4767 4772 4777 4782 4787 4792 4797 4802 4807 4812 4817 4822 4827 4832 4837 4842 4847 4852 4857 4862 4867 4872 4877 4882 4887 4892 4897 4902 4907 4912 4917 4922 4927 4932 4937 4942 4947 4952 4957 4962 4967 4972 4977 4982 4987 4992 4997 5002 5007 5012 5017 5022 5027 5032 5037 5042 5047 5052 5057 5062 5067 5072 5077 5082 5087 5092 5097 5102 5107 5112 5117 5122 5127 5132 5137 5142 5147 5152 5157 5162 5167 5172 5177 5182 5187 5192 5197 5202 5207 5212 5217 5222 5227 5232 5237 5242 5247 5252 5257 5262 5267 5272 5277 5282 5287 5292 5297 5302 5307 5312 5317 5322 5327 5332 5337 5342 5347 5352 5357 5362 5367 5372 5377 5382 5387 5392 5397 5402 5407 5412 5417 5422 5427 5432 5437 5442 5447 5452 5457 5462 5467 5472 5477 5482 5487 5492 5497 5502 5507 5512 5517 5522 5527 5532 5537 5542 5547 5552 5557 5562 5567 5572 5577 5582 5587 5592 5597 5602 5607 5612 5617 5622 5627 5632 5637 5642 5647 5652 5657 5662 5667 5672 5677 5682 5687 5692 5697 5702 5707 5712 5717 5722 5727 5732 5737 5742 5747 5752 5757 5762 5767 5772 5777 5782 5787 5792 5797 5802 5807 5812 5817 5822 5827 5832 5837 5842 5847 5852 5857 5862 5867 5872 5877 5882 5887 5892 5897 5902 5907 5912 5917 5922 5927 5932 5937 5942 5947 5952 5957 5962 5967 5972 5977 5982 5987 5992 5997 6002 6007 6012 6017 6022 6027 6032 6037 6042 6047 6052 6057 6062 6067 6072 6077 6082 6087 6092 6097 6102 6107 6112 6117 6122 6127 6132 6137 6142 6147 6152 6157 6162 6167 6172 6177 6182 6187 6192 6197 6202 6207 6212 6217 6222 6227 6232 6237 6242 6247 6252 6257 6262 6267 6272 6277 6282 6287 6292 6297 6302 6307 6312 6317 6322 6327 6332 6337 6342 6347 6352 6357 6362 6367 6372 6377 6382 6387 6392 6397 6402 6407 6412 6417 6422 6427 6432 6437 6442 6447 6452 6457 6462 6467 6472 6477 6482 6487 6492 6497 6502 6507 6512 6517 6522 6527 6532 6537 6542 6547 6552 6557 6562 6567 6572 6577 6582 6587 6592 6597 6602 6607 6612 6617 6622 6627 6632 6637 6642 6647 6652 6657 6662 6667 6672 6677 6682 6687 6692 6697 6702 6707 6712 6717 6722 6727 6732 6737 6742 6747 6752 6757 6762 6767 6772 6777 6782 6787 6792 6797 6802 6807 6812 6817 6822 6827 6832 6837 6842 6847 6852 6857 6862 6867 6872 6877 6882 6887 6892 6897 6902 6907 6912 6917 6922 6927 6932 6937 6942 6947 6952 6957 6962 6967 6972 6977 6982 6987 6992 6997 7002 7007 7012 7017 7022 7027 7032 7037 7042 7047 7052 7057 7062 7067 7072 7077 7082 7087 7092 7097 7102 7107 7112 7117 7122 7127 7132 7137 7142 7147 7152 7157 7162 7167 7172 7177 7182 7187 7192 7197 7202 7207 7212 7217 7222 7227 7232 7237 7242 7247 7252 7257 7262 7267 7272 7277 7282 7287 7292 7297 7302 7307 7312 7317 7322 7327 7332 7337 7342 7347 7352 7357 7362 7367 7372 7377 7382 7387 7392 7397 7402 7407 7412 7417 7422 7427 7432 7437 7442 7447 7452 7457 7462 7467 7472 7477 7482 7487 7492 7497 7502 7507 7512 7517 7522 7527 7532 7537 7542 7547 7552 7557 7562 7567 7572 7577 7582 7587 7592 7597 7602 7607 7612 7617 7622 7627 7632 7637 7642 7647 7652 7657 7662 7667 7672 7677 7682 7687 7692 7697 7702 7707 7712 7717 7722 7727 7732 7737 7742 7747 7752 7757 7762 7767 7772 7777 7782 7787 7792 7797 7802 7807 7812 7817 7822 7827 7832 7837 7842 7847 7852 7857 7862 7867 7872 7877 7882 7887 7892 7897 7902 7907 7912 7917 7922 7927 7932 7937 7942 7947 7952 7957 7962 7967 7972 7977 7982 7987 7992 7997 8002 8007 8012 8017 8022 8027 8032 8037 8042 8047 8052 8057 8062 8067 8072 8077 8082 8087 8092 8097 8102 8107 8112 8117 8122 8127 8132 8137 8142 8147 8152 8157 8162 8167 8172 8177 8182 8187 8192 8197 8202 8207 8212 8217 8222 8227 8232 8237 8242 8247 8252 8257 8262 8267 8272 8277 8282 8287 8292 8297 8302 8307 8312 8317 8322 8327 8332 8337 8342 8347 8352 8357 8362 8367 8372 8377 8382 8387 8392 8397 8402 8407 8412 8417 8422 8427 8432 8437 8442 8447 8452 8457 8462 8467 8472 8477 8482 8487 8492 8497 8502 8507 8512 8517 8522 8527 8532 8537 8542 8547 8552 8557 8562 8567 8572 8577 8582 8587 8592 8597 8602 8607 8612 8617 8622 8627 8632 8637 8642 8647 8652 8657 8662 8667 8672 8677 8682 8687 8692 8697 8702 8707 8712 8717 8722 8727 8732 8737 8742 8747 8752 8757 8762 8767 8772 8777 8782 8787 8792 8797 8802 8807 8812 8817 8822 8827 8832 8837 8842 8847 8852 8857 8862 8867 8872 8877 8882 8887 8892 8897 8902 8907 8912 8917 8922 8927 8932 8937 8942 8947 8952 8957 8962 8967 8972 8977 8982 8987 8992 8997 9002 9007 9012 9017 9022 9027 9032 9037 9042 9047 9052 9057 9062 9067 9072 9077 9082 9087 9092 9097 9102 9107 9112 9117 9122 9127 9132 9137 9142 9147 9152 9157 9162 9167 9172 9177 9182 9187 9192 9197 9202 9207 9212 9217 9222 9227 9232 9237 9242 9247 9252 9257 9262 9267 9272 9277 9282 9287 9292 9297 9302 9307 9312 9317 9322 9327 9332 9337 9342 9347 9352 9357 9362 9367 9372 9377 9382 9387 9392 9397 9402 9407 9412 9417 9422 9427 9432 9437 9442 9447 9452 9457 9462 9467 9472 9477 9482 9487 9492 9497 9502 9507 9512 9517 9522 9527 9532 9537 9542 9547 9552 9557 9562 9567 9572 9577 9582 9587 9592 9597 9602 9607 9612 9617 9622 9627 9632 9637 9642 9647 9652 9657 9662 9667 9672 9677 9682 9687 9692 9697 9702 9707 9712 9717 9722 9727 9732 9737 9742 9747 9752 9757 9762 9767 9772 9777 9782 9787 9792 9797 9802 9807 9812 9817 9822 9827 9832 9837 9842 9847 9852 9857 9862 9867 9872 9877 9882 9887 9892 9897 9902 9907 9912 9917 9922 9927 9932 9937 9942 9947 9952 9957 9962 9967 9972 9977 9982 9987 9992 9997 10002 10007 10012 10017 10022 10027 10032 10037 10042 10047 10052 10057 10062 10067 10072 10077 10082 10087 10092 10097 10102 10107 10112 10117 10122 10127 10132 10137 10142 10147 10152 10157 10162 10167 10172 10177 10182 10187 10192 10197 10202 10207 10212 10217 10222 10227 10232 10237 10242 10247 10252 10257 10262 10267 10272 10277 10282 10287 10292 10297 10302 10307 10312 10317 10322 10327 10332 10337 10342 10347 10352 10357 10362 10367 10372 10377 10382 10387 10392 10397 10402 10407 10412 10417 10422 10427 10432 10437 10442 10447 10452 10457 10462 10467 10472 10477 10482 10487 10492 10497 10502 10507 10512 10517 10522 10527 10532 10537 10542 10547 10552 10557 10562 10567 10572 10577 10582 10587 10592 10597 10602 10607 10612 10617 10622 10627 10632 10637 10642 10647 10652 10657 10662 10667 10672 10677 10682 10687 10692 10697 10702 10707 10712 10717 10722 10727 10732 10737 10742 10747 10752 10757 10762 10767 10772 10777 10782 10787 10792 10797 10802 10807 10812 10817 10822 10827 10832 10837 10842 10847 10852 10857 10862 10867 10872 10877 10882 10887 10892 10897 10902 10907 10912 10917 10922 10927 10932 10937 10942 10947 10952 10957 10962 10967 10972 10977 10982

21 mg of tetrazole  
1 ml anhydrous  $\text{CH}_3\text{CN}$

5 [0276] After being treated for approximately 8 min, the slide was washed off with  $\text{CH}_3\text{CN}$ , dried, and oxidized with  $\text{I}_2/\text{H}_2\text{O}/\text{THF}/\text{lutidine}$  for 1 min. The slide was again washed, dried, and treated for 30 min with a 20% solution of DBU in DMF. After thorough rinsing of the slide, it was next exposed to a FITC solution (1mM fluorescein isothiocyanate [FITC] in DMF) for 50 min, then washed, dried, and examined by fluorescence microscopy. This reaction is illustrated:



40 3. Preparation and Detection of Thymidine-Cytidine dimer (Biotin)

[0277] An aminopropyl glass slide, was soaked in a solution of ethylene oxide (20% in DMF) to generate a hydroxylated surface. The slide was added a mixture of the following:

45 32 mg of NVO-T-OCED (X)  
11 mg of tetrazole  
0.5 ml of anhydrous  $\text{CH}_3\text{CN}$

50 [0278] After 8 min the plate was then rinsed with acetonitrile, then oxidized with  $\text{I}_2/\text{H}_2\text{O}/\text{THF}/\text{lutidine}$  for 1 min, washed and dried. The slide was then exposed to a 1:3 mixture of acetic anhydride:pyridine for 1 h, then washed and dried. The substrate was then photolyzed in dioxane at 362 nm at 14 mW/cm<sup>2</sup> for 10 min using a 500 $\mu\text{m}$  checkerboard mask, dried, and then treated with a mixture of the following:

55 65 mg of biotin modified C (IV)  
11 mg of tetrazole  
0.5 ml anhydrous  $\text{CH}_3\text{CN}$

[0279] After 8 min the slide was washed with  $\text{CH}_3\text{CN}$  then oxidized with  $\text{I}_2/\text{H}_2\text{O}/\text{THF}/\text{lutidine}$  for 1 min, washed, and

then dried. The slide was then soaked for 30 min in a PBS/0.05% Tween 20 buffer and the solution then shaken off. The slide was next treated with FITC-labeled streptavidin at 10  $\mu$ g/ml in the same buffer system for 30 min. After this time the streptavidin-buffer system was rinsed off with fresh PBS/0.05% Tween 20 buffer and then the slide was finally agitated in distilled water for about 1/2 h. After drying, the slide was examined by fluorescence microscopy (see Fig. 5 and Fig. 3).

#### 4. substrate preparation

[0280] Before attachment of reactive groups it is preferred to clean the substrate which is, in a preferred embodiment, a glass substrate such as a microscope slide or cover slip. A roughened surface will be useable but a plastic or other solid substrate is also appropriate. According to one embodiment the slide is soaked in an alkaline bath consisting of, e.g., 1 liter of 95% ethanol with 120 ml of water and 120 grams of sodium hydroxide for 12 hours. The slides are washed with a buffer and under running water, allowed to air dry, and rinsed with a solution of 95% ethanol.

[0281] The slides are then aminated with, e.g., aminopropyltriethoxysilane for the purpose of attaching amino groups to the glass surface on linker molecules, although other omega functionalized silanes could also be used for this purpose. In one embodiment 0.1% aminopropyltriethoxysilane is utilized, although solutions with concentrations from 10-7% to 10% may be used, with about 10-3% to 2% preferred. A 0.1% mixture is prepared by adding to 100 ml of a 95% ethanol/5% water mixture, 100 microliters ( $\mu$ l) of aminopropyltriethoxysilane. The mixture is agitated at about ambient temperature on a rotary shaker for an appropriate amount of time, e.g., about 5 minutes. 500  $\mu$ l of this mixture is then applied to the surface of one side of each cleaned slide. After 4 minutes or more, the slides are decanted of this solution and thoroughly rinsed three times or more by dipping in 100% ethanol.

[0282] After the slides dry, they are heated in a 110-120°C vacuum oven for about 20 minutes, and then allowed to cure at room temperature for about 12 hours in an argon environment. The slides are then dipped into DMF (dimethylformamide) solution, followed by a thorough washing with methylene chloride.

#### 5. linker attachment, blocking of free sites

[0283] The aminated surface of the slide is then exposed to about 500  $\mu$ l of, for example, a 30 millimolar (mM) solution of NVOC-nucleotide- NHS (N-hydroxysuccinimide) in DMF for attachment of a NVOC-nucleotide to each of the amino groups. See, e.g., SIGMA Chemical Company for various nucleotide derivatives. The surface is washed with, for example, DMF, methylene chloride, and ethanol.

[0284] Any unreacted aminopropyl silane on the surface, i.e., those amino groups which have not had the NVOC-nucleotide attached, are now capped with acetyl groups (to prevent further reaction) by exposure to a 1:3 mixture of acetic anhydride in pyridine for 1 hour. Other materials which may perform this residual capping function include trifluoroacetic anhydride, formicacetic anhydride, or other reactive acylating agents. Finally, the slides are washed again with DMF, methylene chloride, and ethanol.

#### 6. synthesis of eight trimers of C and T

[0285] Fig. 4 illustrates a possible synthesis of the eight trimers of the two-monomer set: cytosine and thymine (represented by C and T, respectively). A glass slide bearing silane groups terminating in 6-nitroveratryloxycarboxamide (NVOC-NH) residues is prepared as a substrate. Active esters (pentafluorophenyl, OBT, etc.) of cytosine and thymine protected at the 5' hydroxyl group with NVOC are prepared as reagents. While not pertinent to this example, if side chain protecting groups are required for the monomer set, these must not be photoreactive at the wavelength of light used to protect the primary chain.

[0286] For a monomer set of size  $n$ ,  $n \times \ell$  cycles are required to synthesize all possible sequences of length  $\ell$ . A cycle consists of:

1. Irradiation through an appropriate mask to expose the 5'-OH groups at the sites where the next residue is to be added, with appropriate washes to remove the by-products of the deprotection.
2. Addition of a single activated and protected (with the same photochemically-removable group) monomer, which will react only at the sites addressed in step 1, with appropriate washes to remove the excess reagent from the surface.

[0287] The above cycle is repeated for each member of the monomer set until each location on the surface has been extended by one residue in one embodiment. In other embodiments, several residues are sequentially added at one location before moving on to the next location. Cycle times will generally be limited by the coupling reaction rate, now as short as about 10 min in automated oligonucleotide synthesizers. This step is optionally followed by addition of a

protecting group to stabilize the array for later testing. For some types of polymers (e.g., peptides), a final deprotection of the entire surface (removal of photoprotective side chain groups) may be required.

[0288] More particularly, as shown in Fig. 4A, the glass 20 is provided with regions 22, 24, 26, 28, 30, 32, 34, and 36. Regions 30, 32, 34, and 36 are masked, indicated by the hatched regions, as shown in Fig. 4B and the glass is irradiated by the bright regions 22, 24, 26, and 28, and exposed to a reagent containing a photosensitive blocked C (e.g., cytosine derivative), with the resulting structure shown in Fig. 4C. The substrate is carefully washed and the reactants removed. Thereafter, regions 22, 24, 26, and 28 are masked, as indicated by the hatched region, the glass is irradiated (as shown in Fig. 4D), as indicated by the bright regions, at 30, 32, 34, and 36, and exposed to a photo-sensitive blocked reagent containing T (e.g., thymine derivative), with the resulting structure shown in Fig. 4E. The process proceeds, consecutively masking and exposing the sections as shown until the structure shown in Fig. 4M is obtained. The glass is irradiated and the terminal groups are, optionally, capped by acetylation. As shown, all possible trimers of cytosine/thymine are obtained.

[0289] In this example, no side chain protective group removal is necessary, as might be common in modified nucleotides. If it is desired, side chain deprotection may be accomplished by treatment with ethanedithiol and trifluoroacetic acid.

[0290] In general, the number of steps needed to obtain a particular polymer chain is defined by:

$$n \times \ell \quad (1)$$

20 where:

25  $n$  = the number of monomers in the basis set of monomers, and  
 $\ell$  = the number of monomer units in a polymer chain.

[0291] Conversely, the synthesized number of sequences of length  $\ell$  will be:

$$n^\ell \quad (2)$$

30 [0292] Of course, greater diversity is obtained by using masking strategies which will also include the synthesis of polymers having a length of less than  $\ell$ . If, in the extreme case, all polymers having a length less than or equal to  $\ell$  are synthesized, the number of polymers synthesized will be:

$$n^\ell + n^{\ell-1} + \dots + n^1 \quad (3)$$

40 [0293] The maximum number of lithographic steps needed will generally be  $n$  for each "layer" of monomers, i.e., the total number of masks (and, therefore, the number of lithographic steps) needed will be  $n \times \ell$ . The size of the transparent mask regions will vary in accordance with the area of the substrate available for synthesis and the number of sequences to be formed. In general, the size of the synthesis areas will be:

$$45 \text{size of synthesis areas} = (A)/(S)$$

where:

50 A is the total area available for synthesis; and  
S is the number of sequences desired in the area.

[0294] It will be appreciated by those of skill in the art that the above method could readily be used to simultaneously produce thousands or millions of oligomers on a substrate using the photolithographic techniques disclosed herein. Consequently, the method results in the ability to practically test large numbers of, for example, di, tri, tetra, penta, hexa, hepta, octa, nona, deca; even dodecanucleotides, or larger polynucleotides.

55 [0295] The above example has illustrated the method by way of a manual example. It will of course be appreciated that automated or semi-automated methods could be used. The substrate would be mounted in a flow cell for automated addition and removal of reagents, to minimize the volume of reagents needed, and to more carefully control reaction

conditions. Successive masks will be applicable manually or automatically. See, e.g., PCT publication no. WO90/15070.

7. labeling of target

5 [0296] The target oligonucleotide can be labeled using standard procedures referred to above. As discussed, for certain situations, a reagent which recognizes interaction, e.g., ethidium bromide, may be provided in the detection step. Alternatively, fluorescence labeling techniques may be applied, see, e.g., Smith, et al. (1986) Nature, 321: 674-679; and Prober, et al. (1987) Science, 238:336-341. The techniques described therein will be followed with minimal modifications as appropriate for the label selected.

10 8. dimers of A, C, G, and T

15 [0297] The described technique may be applied, with photosensitive blocked nucleotides corresponding to adenine, cytosine, guanine, and thymine, to make combinations of polynucleotides consisting of each of the four different nucleotides. All 16 possible dimers would be made using a minor modification of the described method.

19 9. 10-mers of A, C, G, and T

20 [0298] The described technique for making dimers of A, C, G, and T may be further extended to make longer oligonucleotides. The automated system described, e.g., in PCT publication no. WO90/15070 can be adapted to make all possible 10-mers composed of the 4 nucleotides A, C, G, and T. The photosensitive, blocked nucleotide analogues have been described above, and would be readily adaptable to longer oligonucleotides.

25 10. specific recognition hybridization to 10-mers

30 [0299] The described hybridization conditions are directly applicable to the sequence specific recognition reagents attached to the substrate, produced as described immediately above. The 10-mers have an inherent property of hybridizing to a complementary sequence. For optimum discrimination between full matching and some mismatch, the conditions of hybridization should be carefully selected, as described above. Careful control of the conditions, and titration of parameters should be performed to determine the optimum collective conditions.

35 11. hybridization

40 [0300] Hybridization conditions are described in detail, e.g., in Hames and Higgins (1985) Nucleic Acid Hybridisation: A Practical Approach; and the considerations for selecting particular conditions are described, e.g., in Wetmur and Davidson, (1988) J. Mol. Biol. 31:349-370, and Wood et al. (1985) Proc. Natl. Acad. Sci. USA 82:1585-1588. As described above, conditions are desired which can distinguish matching along the entire length of the probe from where there is one or more mismatched bases. The length of incubation and conditions will be similar, in many respects, to the hybridization conditions used in Southern blot transfers. Typically, the GC bias may be minimized by the introduction of appropriate concentrations of the alkylammonium buffers, as described above.

45 [0301] Titration of the temperature and other parameters is desired to determine the optimum conditions for specificity and distinguishability of absolutely matched hybridization from mismatched hybridization.

[0302] A fluorescently labeled target or set of targets are generated, as described in Prober, et al. (1987) Science 238:336-341, or Smith, et al. (1986) Nature 321:674-679. Preferably, the target or targets are of the same length as, or slightly longer, than the oligonucleotide probes attached to the substrate and they will have known sequences. Thus, only a few of the probes hybridize perfectly with the target, and which particular ones did would be known.

50 [0303] The substrate and probes are incubated under appropriate conditions for a sufficient period of time to allow hybridization to completion. The time is measured to determine when the probe-target hybridizations have reached completion. A salt buffer which minimizes GC bias is preferred, incorporating, e.g., buffer, such as tetramethyl ammonium or tetraethyl ammonium ion at between about 2.4 and 3.0 M. See Wood, et al. (1985) Proc. Nat'l Acad. Sci. USA 82:1585-1588. This time is typically at least about 30 min, and may be as long as about 1-5 days. Typically very long matches will hybridize more quickly, very short matches will hybridize less quickly, depending upon relative target and probe concentrations. The hybridization will be performed under conditions where the reagents are stable for that time duration.

55 [0304] Upon maximal hybridization, the conditions for washing are titrated. Three parameters initially titrated are time, temperature, and cation concentration of the wash step. The matrix is scanned at various times to determine the conditions at which the distinguishability between true perfect hybrid and mismatched hybrid is optimized. These conditions will be preferred in the sequencing embodiments.

## 12. positional detection of specific interaction

[0305] As indicated above, the detection of specific interactions may be performed by detecting the positions where the labeled target sequences are attached. Where the label is a fluorescent label, the apparatus described, e.g., PCT publication no. WO90/15070 may be advantageously applied. In particular, the synthetic processes described above will result in a matrix pattern of specific sequences attached to the substrate, and a known pattern of interactions can be converted to corresponding sequences.

[0306] In an alternative embodiment, a separate reagent which differentially interacts with the probe and interacted probe targets can indicate where interaction occurs or does not occur. A single-strand specific reagent will indicate where no interaction has taken place, while a double-strand specific reagent will indicate where interaction has taken place. An intercalating dye, e.g., ethidium bromide, may be used to indicate the positions of specific interaction.

## 13. analysis

[0307] Conversion of the positional data into sequence specificity will provide the set of subsequences whose analysis by overlap segments, may be performed, as described above. Analysis is provided by the methodology described above, or using, e.g., software available from the Genetic Engineering Center, P.O. Box 794, 11000 Belgrade, Yugoslavia (Yugoslav group). See, also, Macevitz, PCT publication no. WO 90/04652.

[0308] Preparation of short peptides on a substrate is described below.

## POLYNUCLEOTIDE FINGERPRINTING

[0309] The above section on generation of reagents for sequencing provides specific reagents useful for fingerprinting applications. Fingerprinting embodiments may be applied towards polynucleotide fingerprinting, cell and tissue classification, cell and tissue temporal development stage classification, diagnostic tests, forensic uses for individual identification, classification of organisms, and genetic screening of individuals. Mapping applications are also described below.

[0310] Polynucleotide fingerprinting may use reagents similar to those described above for probing a sequence for the presence of specific subsequences found therein. Typically, the subsequences used for fingerprinting will be longer than the sequences used in oligonucleotide sequencing. In particular, specific long segments may be used to determine the similarity of different samples of nucleic acids. They may also be used to fingerprint whether specific combinations of information are provided therein. Particular probe sequences are selected and attached in a positional manner to a substrate. The means for attachment may be either using a caged biotin method described or by another method using targeting molecules. In one embodiment, an unnatural nucleotide or similar complementary binding molecule may be attached to the fingerprinting probe and the probe thereby directed towards complementary sequences on a VLSIPS substrate. Typically, unnatural nucleotides would be preferred, e.g., unnatural optical isomers, which would not interfere with natural nucleotide interactions.

[0311] Having produced a substrate with particular fingerprint probes attached thereto at positionally defined regions, the substrate may be used in a manner quite similar to the sequencing embodiment to provide information as to whether the fingerprint probes are detecting the corresponding sequence in a target sequence. This will often provide information similar to a Southern blot hybridization.

Temporal Development

## 45 Developmental RNA expression patterns

[0312] The present fingerprinting invention also allows cell classification by identification of developmental RNA expression patterns. For example, a lymphocyte stem cell expresses a particular combination of RNA species. As the lymphocyte develops through a program developmental scheme, at various stages it expresses particular RNA species which are diagnostic of particular stages in development. Again, the fingerprinting methodology allows for the definition of specific structural features which are diagnostic of developmental or functional features which will allow classification of cells into temporal developmental classes. Cells, products of those cells, or lysates of those cells will be assayed to determine the developmental stage of the source cells. In this manner, once a developmental stage is defined, specific synchronized populations of cells will be selected out of another population. These synchronized populations may be very important in determining the biological mechanisms of development.

[0313] The present invention also allows for fingerprinting of the mRNA population of a cell. In this fashion, the mRNA population, which should be a good determinant of developmental stage, will be correlated with other structural features of the cell. In this manner, cells at specific developmental stages will be characterized by the intracellular environment,

as well as the extracellular environment.

#### Diagnostic Tests

5 [0314] The present invention also provides the ability to perform diagnostic tests. Diagnostic tests typically are based upon a fingerprint type assay, which tests for the presence of specific diagnostic polynucleotides. Thus, the present invention provides means for viral strain identification, bacterial strain identification, and other diagnostic tests using positionally defined specific oligonucleotide reagents.

10 Viral Identification

[0315] The present invention provides reagents and methodology for identifying viral strains. The viral genome may be probed for specific sequences which are characteristic of particular viral strains. Specific hybridization patterns on an VLSIPS oligonucleotide substrate can identify the presence of particular viral genomes.

15 Bacterial Identification

20 [0316] Similar techniques will be applicable to identifying a bacterial source. This may be useful in diagnosing bacterial infections, or in classifying sources of particular bacterial species. For example, the bacterial assay may be useful in determining the natural range of survivability of particular strains of bacteria across regions of the country or in different ecological niches.

Other Microbiological Identifications

25 [0317] The present invention provides means for diagnosis of other microbiological and other species, e.g., protozoal species and parasitic species in a biological sample, but also provides the means for assaying a combination of different infections. For example, a biological specimen may be assayed for the presence of any or all of these microbiological species. In human diagnostic uses, typical samples will be blood, sputum, stool, urine, or other samples.

30 Individual Identification

35 [0318] The present invention provides the ability to fingerprint and identify a genetic individual. This individual may be a bacterial or lower microorganism, as described above in diagnostic tests, or of a plant or animal. An individual may be identified genetically, as described.

[0319] Genetic fingerprinting has been utilized in comparing different related species in Southern hybridization blots. Genetic fingerprinting has also been used in forensic studies, see, e.g., Morris et al. (1989) *J. Forensic Science* 34: 1311-1317, and references cited therein. As described above, an individual may be identified genetically by a sufficiently large number of probes. The likelihood that another individual would have an identical pattern over a sufficiently large number of probes may be statistically negligible. However, it is often quite important that a large number of probes be used where the statistical probability of matching is desired to be particularly low. In fact, the probes will optimally be selected for having high heterogeneity among the population. In addition, the fingerprint method may make use of the pattern of homologies indicated by a series of more and more stringent washes. Then, each position has both a sequence specificity and a homology measurement, the combination of which greatly increases the number of dimensions and the statistical likelihood of a perfect pattern match with another genetic individual.

45 Genetic Screening

50 1. test alleles with markers

[0320] The present invention provides for the ability to screen for genetic variations of individuals. For example, a number of genetic diseases are linked with specific alleles. See, e.g., Scriber, C. et al. (eds.) (1989) The Metabolic Bases of Inherited Disease, McGraw-Hill, New York. In one embodiment, cystic fibrosis has been correlated with a specific gene, see, Gregory et al. (1990) *Nature* 347: 382-386. A number of alleles are correlated with specific genetic deficiencies. See, e.g., McKusick, V. (1990) Genetic Inheritance in Man: Catalogs of Autosomal Dominant, Autosomal Recessive, and X-linked Phenotypes, Johns Hopkins University Press, Baltimore; Ott, J. (1985) Analysis of Human Genetic Linkage, Johns Hopkins University Press, Baltimore; Track, R. et al. (1989) Banbury Report 32: DNA Technology and Forensic Science, Cold Spring Harbor Press, New York.

## 2. Amniocentesis

[0321] Typically, amniocentesis is used to determine whether chromosome translocations have occurred. The mapping procedure may provide the means for determining whether these translocations have occurred, and for detecting particular alleles of various markers.

## MAPPING

Positionally Located Clones

[0322] The present invention allows for the positional location of specific clones useful for mapping: For example, caged biotin may be used for specifically positioning a probe to a location on a matrix pattern.

[0323] In addition, the specific probes may be positionally directed to specific locations on a substrate by targeting. For example, polypeptide specific recognition reagents may be attached to oligonucleotide sequences which can be complementarily targeted, by hybridization, to specific locations on a VLSIPS substrate. Hybridization conditions, as applied for oligonucleotide probes, will be used to target the reagents to locations on a substrate having complementary oligonucleotides synthesized thereon. In another embodiment, oligonucleotide probes may be attached to specific polypeptide targeting reagents such as an antigen or antibody. These reagents can be directed towards a complementary antigen or antibody already attached to a VLSIPS substrate.

[0324] In another embodiment, an unnatural nucleotide which does not interfere with natural nucleotide complementary hybridization may be used to target oligonucleotides to particular positions on a substrate. Unnatural optical isomers of natural nucleotides should be ideal candidates.

[0325] In this way, short probes may be used to determine the mapping of long targets or long targets may be used to map the position of shorter probes. See, e.g., Craig et al. 1990 *Nuc. Acids Res.* 18: 2653-2660.

Positionally Defined Clones

[0326] Positionally defined clones may be transferred to a new substrate by either physical transfer or by synthetic means. Synthetic means may involve either a production of the probe on the substrate using the VLSIPS synthetic methods, or may involve the attachment of a targeting sequence made by VLSIPS synthetic methods which will target that positionally defined clone to a position on a new substrate. Both methods will provide a substrate having a number of positionally defined probes useful in mapping.

## CONCLUSION

[0327] The present inventions provide greatly improved methods and apparatus for synthesis of polymers on substrates. It is to be understood that the above description is intended to be illustrative and not restrictive. Many embodiments will be apparent to those of skill in the art upon reviewing the above description. By way of example, the invention has been described primarily with reference to the use of photoremovable protective groups, but it will be readily recognized by those of skill in the art that sources of radiation other than light could also be used. For example, in some embodiments it may be desirable to use protective groups which are sensitive to electron beam irradiation, x-ray irradiation, in combination with electron beam lithograph, or x-ray lithography techniques. Alternatively, the group could be removed by exposure to an electric current. The scope of the invention should, therefore, be determined not with reference to the above description, but should instead be determined with reference to the appended claims, along with the full scope of equivalents to which such claims are entitled.

## Claims

50 1. A method for detecting nucleic acid sequences in two or more collections of nucleic acids, comprising:

(a) providing an array comprising more than 100 different polynucleotide probes bound to a solid surface;

(b) contacting said array of probes under hybridisation conditions with:

55 (i) a first collection of nucleic acids comprised of first-labelled nucleic acids having at least some sequences complementary to probes of said array, and

5 (ii) at least a second collection of nucleic acids comprised of second-labelled nucleic acids having at least some sequences complementary to probes of said array,

wherein said first and second labels are distinguishable from each other; and

10 (c) detecting hybridisation of first and second labelled complementary nucleic acids to probes of said array.

15 2. A method as claimed in claim 1, wherein said first and second labels are fluorescent labels that emit light of different wavelengths.

20 3. A method as claimed in claim 2 used to fingerprint at least first and second cells, wherein said first collection of nucleic acids is from a first cell and said second collection of nucleic acids is from a second cell, and fluorescence of said first and second labels hybridised to the array is detected, optionally said method further comprising:

25 (a) determining levels of gene expression in said first and second cells,

(b) determining patterns of gene expression in said first and second cells, or

30 (c) determining genetic differences between said first and second cells.

20 4. A method as claimed in claim 3, wherein said first and second cells are different types of cells, optionally wherein:

25 (a) at least one cell type is a tumour cell or other cell exhibiting abnormal physiology,

30 (b) said first and second cells are at different stages of development,

(c) said first and second cells are at different stages of infection or other disease, or

35 (d) said first and second cells are from different species of organism, optionally wherein said organism is an animal, plant or microorganism.

5. A method as claimed in claim 3 or claim 4, wherein at least one collection of nucleic acids is synthesized by fluorescently labelling:

35 (a) RNA isolated, generated or amplified from said cell; or

40 (b) DNA isolated, generated or amplified from said cell.

6. A method as claimed in any one of claims 1 to 5, wherein said solid surface is a polymeric substrate or includes fibers.

45 7. A method as claimed in any one of claims 1 to 6, wherein said probes are bound at a density of at least  $10^3$ , preferably at least  $10^4$ , more preferably at least  $10^5$ , even more preferably at least  $10^6$  regions per  $\text{cm}^2$  to known regions on the solid surface.

8. A method as claimed in any one of claims 1 to 5, wherein said solid surface is formed as a collection of beads and each different polynucleotide probe is bound to a single bead.

50 9. A method as claimed in claim 8, wherein a bead further has an encoding system bound thereto such that the sequence of the polynucleotide bound to a bead can be determined by decoding the encoding system, optionally wherein said encoding system is selected from the group consisting of a magnetic system, shape encoding system, colour encoding system, or combination thereof.

10. A method as claimed in claim 8 or claim 9, wherein an automated cell sorter is used to detect hybridisation.

55 11. A method as claimed in any one of claims 1 to 10, wherein said array is comprised of more than  $10^3$ , preferably more than  $10^4$ , more preferably more than  $10^5$ , even more preferably more than  $10^6$  different probes bound to the solid surface.

12. A method as claimed in any preceding claim, wherein said probes are greater than about 15, preferably greater than about 25, more preferably greater than about 50 nucleotides in length.

5 13. A method as claimed in any preceding claim, wherein at least said two collections of nucleic acids are hybridised to the same array of said probes.

14. A method as claimed in claim 13, wherein at least said two collections of nucleic acids are hybridised separately or simultaneously to the same array of said probes.

10 15. A method as claimed in any preceding claim, wherein said array has been recycled for use.

16. A method as claimed in any preceding claim, wherein the sequences of the polynucleotide probes of the array are known.

15

### Patentansprüche

1. Ein Verfahren zum Nachweis von Nucleinsäuresequenzen in zwei oder mehr Nucleinsäuregruppen, umfassend:

20 (a) Bereitstellen einer Gruppierung, die mehr als 100 verschiedene Polynukleotidsonden aufweist, die an eine feste Oberfläche gebunden sind;

(b) In-Kontakt-Bringen der Gruppierung von Sonden unter Hybridisierungsbedingungen mit:

25 (i) einer ersten Gruppe von Nucleinsäuren, die mit einem ersten Marker markierte Nucleinsäuren umfasst, die mindestens einige Sequenzen aufweisen, die zu Sonden der Gruppierung komplementär sind, und

(ii) mindestens einer zweiten Gruppe von Nucleinsäuren, die mit einem zweiten Marker markierte Nucleinsäuren umfasst, die mindestens einige Sequenzen aufweisen, die zu Sonden der Gruppierung komplementär sind,

30 wobei der erste und der zweite Marker voneinander unterscheidbar sind; und

(c) Nachweisen der Hybridisierung der mit einem ersten Marker und einem zweiten Marker markierten komplementären Nucleinsäuren an Sonden der Gruppierung.

35 2. Verfahren nach Anspruch 1, wobei der erste und der zweite Marker Fluoreszenzmarker sind, die Licht mit verschiedenen Wellenlängen emittieren.

40 3. Verfahren nach Anspruch 2, das zum Anfertigen eines Fingerprints von mindestens ersten und zweiten Zellen verwendet wird, wobei die erste Gruppe von Nucleinsäuren von einer ersten Zelle und die zweite Gruppe von Nucleinsäuren von einer zweiten Zelle stammt, und wobei die Fluoreszenz des ersten und des zweiten Markers, die an die Gruppierung hybridisiert haben, nachgewiesen wird, wobei das Verfahren gegebenenfalls ferner umfasst:

45 (a) Bestimmen des Niveaus der Genexpression in den ersten und zweiten Zellen,

(b) Bestimmen von Mustern der Genexpression in den ersten und zweiten Zellen, oder

50 (c) Bestimmen von genetischen Unterschieden zwischen den ersten und zweiten Zellen.

4. Verfahren nach Anspruch 3, wobei die ersten und zweiten Zellen verschiedene Zelltypen sind, wobei gegebenenfalls:

55 (a) mindestens ein Zelltyp eine Tumorzelle oder eine andere Zelle mit abnormer Physiologie ist,

(b) sich die ersten und zweiten Zellen in verschiedenen Entwicklungsstadien befinden,

(c) sich die ersten und zweiten Zellen in verschiedenen Stadien einer Infektion oder einer anderen Erkrankung befinden, oder

5 (d) die ersten und zweiten Zellen von verschiedenen Organismusarten stammen,

wobei der Organismus gegebenenfalls ein Tier, eine Pflanze oder ein Mikroorganismus ist.

5. Verfahren nach Anspruch 3 oder 4, wobei mindestens eine Gruppe von Nucleinsäuren durch Fluoreszenzmarkierung von:

10 (a) RNA, die von der Zelle isoliert, erzeugt oder amplifiziert worden ist; oder

(b) DNA, die von der Zelle isoliert, erzeugt oder amplifiziert worden ist, synthetisiert wird.

15 6. Verfahren nach einem der Ansprüche 1 bis 5, wobei die feste Oberfläche ein polymerer Träger ist oder Fasern umfasst.

20 7. Verfahren nach einem der Ansprüche 1 bis 6, wobei die Sonden mit einer Dichte von mindestens  $10^3$ , vorzugsweise mindestens  $10^4$ , mehr bevorzugt mindestens  $10^5$  und insbesondere mindestens  $10^6$  Bereichen pro  $\text{cm}^2$  an bekannte Bereiche der festen Oberfläche gebunden sind.

25 8. Verfahren nach einem der Ansprüche 1 bis 5, wobei die feste Oberfläche als Gruppe von Kugelchen ausgebildet ist und jede unterschiedliche Polynucleotidsonde an ein einzelnes Kugelchen gebunden ist.

30 9. Verfahren nach Anspruch 8, wobei ein Kugelchen ferner ein daran gebundenes Codierungssystem aufweist, derart, dass die Sequenz des an ein Kugelchen gebundenen Polynucleotids durch Decodieren des Codierungssystems bestimmt werden kann, wobei das Codierungssystem gegebenenfalls aus der Gruppe bestehend aus einem magnetischen System, Gestaltcodierungssystem, Farbcodierungssystem oder einer Kombination davon ausgewählt ist.

35 10. Verfahren nach Anspruch 8 oder 9, wobei zum Nachweis der Hybridisierung eine automatisierte Zellsortierungseinrichtung verwendet wird.

40 11. Verfahren nach einem der Ansprüche 1 bis 10, wobei die Gruppierung mehr als  $10^3$ , vorzugsweise mehr als  $10^4$ , mehr bevorzugt mehr als  $10^5$  und insbesondere mehr als  $10^6$  verschiedene, an die feste Oberfläche gebundene Sonden umfasst.

45 12. Verfahren nach einem der vorstehenden Ansprüche, wobei die Sonden eine Länge von mehr als etwa 15, vorzugsweise mehr als etwa 25 und insbesondere mehr als etwa 50 Nucleotiden aufweisen.

50 13. Verfahren nach einem der vorstehenden Ansprüche, wobei mindestens die beiden Gruppen von Nucleinsäuren an die gleiche Gruppierung der Sonden hybridisiert werden.

14. Verfahren nach Anspruch 13, wobei mindestens die beiden Gruppen von Nucleinsäuren getrennt oder gleichzeitig an die gleiche Gruppierung von Sonden hybridisiert werden.

15. Verfahren nach einem der vorstehenden Ansprüche, wobei die Gruppierung zur Verwendung rezykliert worden ist.

16. Verfahren nach einem der vorstehenden Ansprüche, wobei die Sequenzen der Polynucleotidsonden der Gruppierung bekannt sind.

#### Revendications

55 1. Méthode pour détecter les séquences nucléiques dans deux ou plusieurs collections d'acides nucléiques, comprenant :

(a) la mise à disposition d'un réseau comprenant plus de 100 différentes sondes polynucléotidiques liées à

une surface solide ;

(b) la mise en contact dudit réseau de sondes dans des conditions d'hybridation avec:

5 (i) une première collection d'acides nucléiques composée d'acides nucléiques fonctionnalisés par un premier marqueur possédant au moins quelques séquences complémentaires aux sondes dudit réseau, et  
 (ii) au moins une seconde collection d'acides nucléiques composée d'acides nucléiques fonctionnalisés à l'aide d'un second marqueur possédant au moins quelques séquences complémentaires aux sondes dudit réseau,

10 avec lesdits premier et second marqueurs étant distincts l'un de l'autre; et

(c) la détection d'hybridation des acides nucléiques complémentaires fonctionnalisés par les premier et second marqueurs aux sondes dudit réseau.

15 2. Méthode selon la revendication 1, **caractérisée en ce que** les premier et second marqueurs sont des marqueurs fluorescents qui émettent de la lumière à des longueurs d'onde différentes.

20 3. Méthode telle que revendiquée en revendication 2, utilisée pour marquer au moins des cellules dites première et seconde, dans laquelle ladite première collection d'acides nucléiques est issue d'une cellule première et ladite seconde collection d'acides nucléiques est issue d'une cellule seconde, et la fluorescence desdits premier et second marqueurs hybridés au réseau est détectée, ladite méthode comprenant en outre le cas échéant:

25 (a) la détermination du niveau d'expression de gène(s) dans lesdites cellules première et seconde,  
 (b) la détermination des profils de gène(s) exprimé(s) dans lesdites cellules première et seconde, ou  
 (c) la détermination des différences génétiques entre lesdites cellules première et seconde.

30 4. Méthode telle que revendiquée en revendication 3, dans laquelle les cellules première et seconde sont des cellules de types différents, le cas échéant **caractérisée en ce que**:

35 (a) au moins un type de cellule est une cellule tumorale ou une autre cellule manifestant une physiologie anormale,  
 (b) lesdites cellules première et seconde sont à des stades différents de développement,  
 (c) lesdites cellules première et seconde sont à des stades différents d'infection ou autre pathologie, ou  
 (d) lesdites cellules première et seconde sont issues d'espèces différentes d'organisme, le cas échéant **caractérisée en ce que** l'organisme est un animal, une plante ou un microorganisme.

40 5. Méthode telle que revendiquée en revendication 3 ou 4, **caractérisée en ce qu'au moins** une collection d'acides nucléiques est synthétisée par marquage de manière fluorescente:

(a) d'ARN isolé, généré ou amplifié à partir de ladite cellule ; ou  
 (b) d'ADN isolé, généré ou amplifié à partir de ladite cellule.

45 6. Méthode telle que revendiquée dans l'une quelconque des revendications 1 à 5, **caractérisée en ce que** la surface solide est un substrat polymérique ou incorpore des fibres.

50 7. Méthode telle que revendiquée dans l'une quelconque des revendications 1 à 6, **caractérisée en ce que** lesdites sondes sont liées à une densité d'au moins  $10^3$ , de préférence d'au moins  $10^4$ , plus préférentiellement d'au moins  $10^5$ , et encore plus préférentiellement d'au moins  $10^6$  régions par  $\text{cm}^2$  à des régions déterminées de la surface solide.

55 8. Méthode telle que revendiquée dans l'une quelconque des revendications 1 à 5, **caractérisée en ce que** ladite surface solide présente une pluralité de reliefs et **en ce que** chaque sonde polynucléotidique différente est liée à un unique relief.

9. Méthode telle que revendiquée en revendication 8, **caractérisée en ce qu'un** relief possède en outre un système de codage lié à sa surface de telle manière que la séquence du polynucléotide lié au relief peut être déterminée en décodant le système de codage, et **en ce que**, le cas échéant, le système de codage est choisi parmi le groupe consistant en un système magnétique, en un système codant une forme, un système codant une couleur, ou l'une de leur combinaison.

10. Méthode telle que revendiquée en revendication 8 ou 9, **caractérisée en ce qu'un trieur de cellule automatique** est utilisé pour détecter l'hybridation.

5 11. Méthode telle que revendiquée dans l'une quelconque des revendications 1 à 10, **caractérisée en ce que** ledit réseau comprend plus de  $10^3$ , de préférence plus de  $10^4$ , plus préférentiellement plus de  $10^5$ , et encore plus préférentiellement plus de  $10^6$  sondes différentes liées à la surface solide.

10 12. Méthode telle que revendiquée dans l'une quelconque des revendications précédentes, **caractérisée en ce que** lesdites sondes possèdent en longueur plus de 15, de préférence environ plus de 25, et plus préférentiellement environ plus de 50 nucléotides.

15 13. Méthode telle que revendiquée dans l'une quelconque des revendications précédentes, **caractérisée en ce qu'au** moins deux collections d'acides nucléiques sont hybridées au même réseau desdites sondes.

20 14. Méthode telle que revendiquée en revendication 13, **caractérisée en ce qu'au moins deux collections d'acides** nucléiques sont hybridées séparément ou simultanément au même niveau desdites sondes.

15 15. Méthode telle que revendiquée dans l'une quelconque des revendications précédentes, **caractérisée en ce que** ledit réseau a été recyclé pour l'utilisation.

25 16. Méthode telle que revendiquée dans l'une quelconque des revendications précédentes, **caractérisée en ce que** les séquences des sondes polynucléotidiques du réseau sont connues.

25

30

35

40

45

50

55

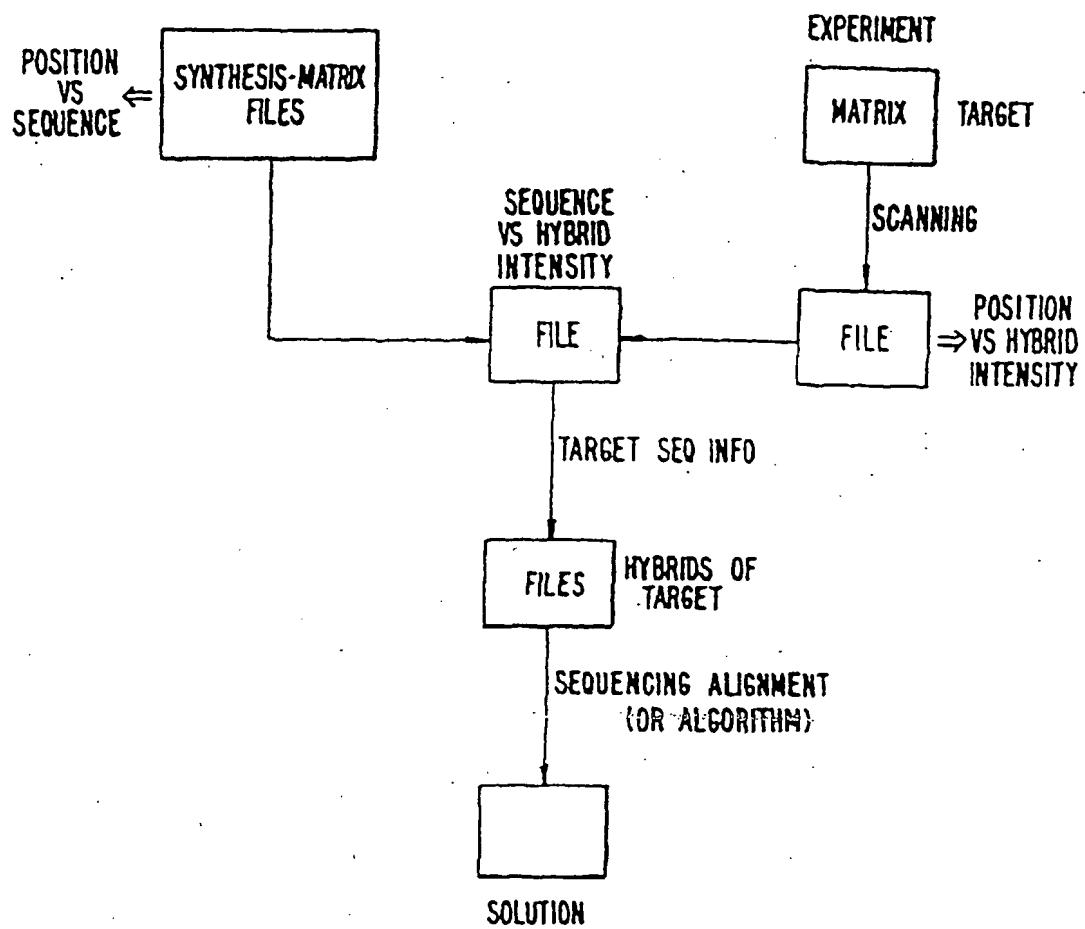


FIG. 1.

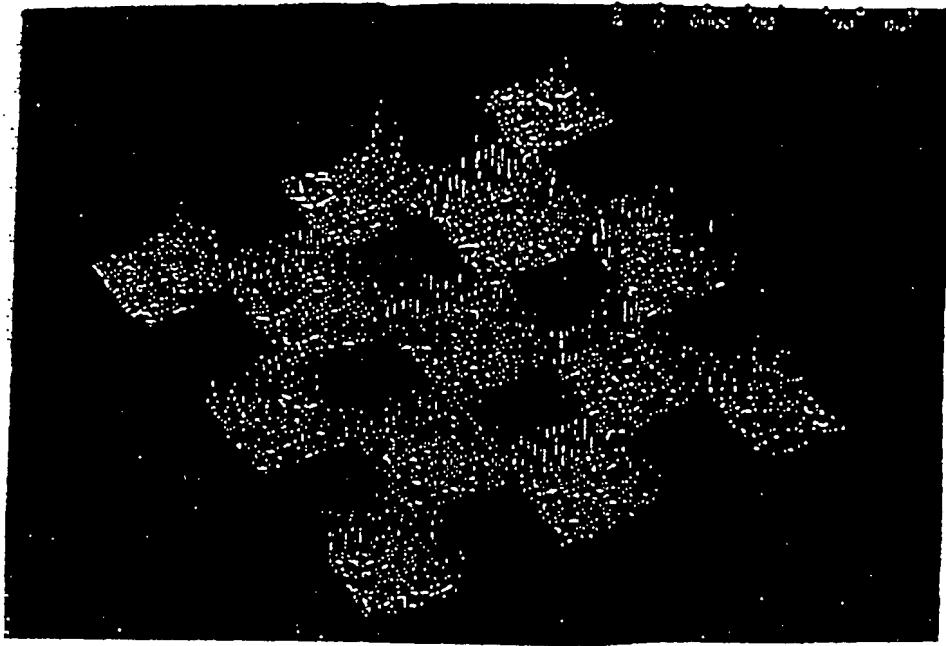


FIG. 2.

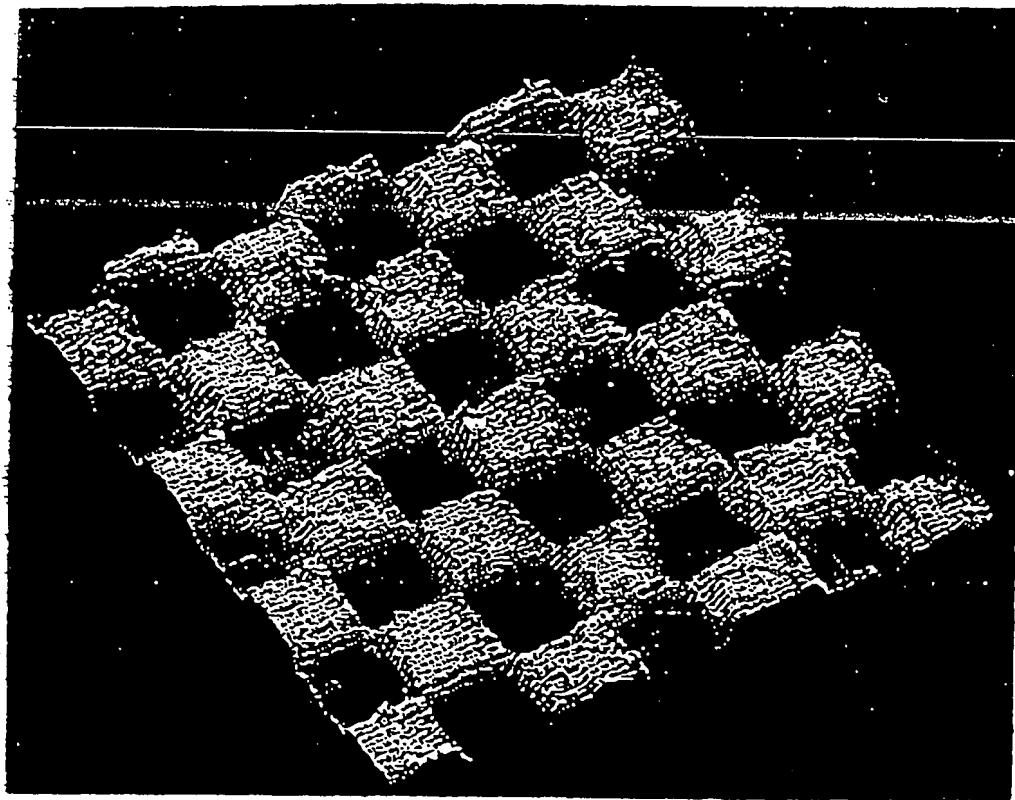


FIG. 3.

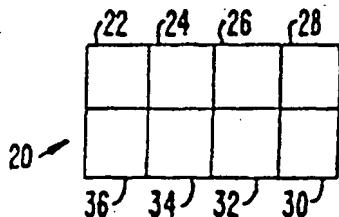


FIG. 4A.

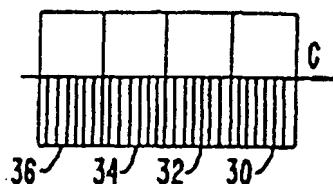


FIG. 4B.

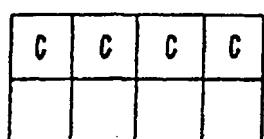


FIG. 4C.

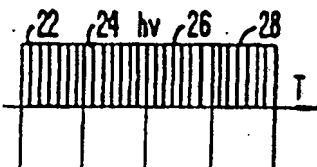


FIG. 4D.

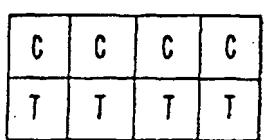


FIG. 4E.

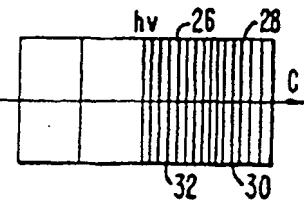


FIG. 4F.

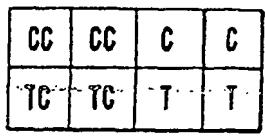


FIG. 4G.

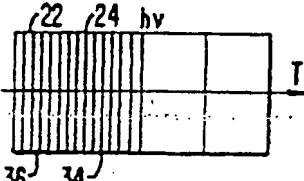


FIG. 4H.

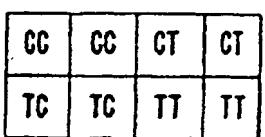


FIG. 4I.

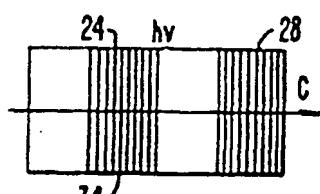


FIG. 4J.

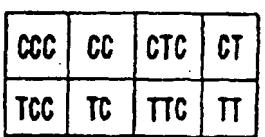


FIG. 4K.

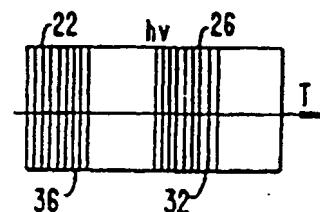


FIG. 4L.

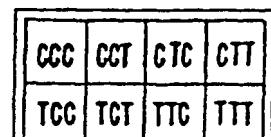


FIG. 4M.